

# Latent Variable Model Selection For Gaussian Conditional Random Fields

Benjamin Frot

Department of Statistics

University Of Oxford

Oxford, United Kingdom

`frot@stats.ox.ac.uk`

Luke Jostins

Wellcome Trust Centre for Human Genetics

University Of Oxford

Oxford, United Kingdom

`lj4@well.ox.ac.uk`

Gil McVean

Department of Statistics &

Wellcome Trust For Human Genetics

University Of Oxford

Oxford, United Kingdom

`mcvean@stats.ox.ac.uk`

## Abstract

We consider the problem of learning a conditional Gaussian graphical model in the presence of latent variables. Building on recent advances in this field, we suggest a method that decomposes the parameters of a conditional Markov random field into the sum of a sparse and a low-rank matrix. We derive convergence bounds for this estimator and show that it is well-behaved in the high-dimensional regime as well as “sparsistent” (i.e. capable of recovering the graph structure). We then describe a proximal gradient algorithm which is able to fit the model to thousands of variables. Through extensive simulations, we illustrate the conditions required for identifiability and show that there is a wide range of situations in which this model performs significantly better than its counterparts. We also show how this problem is relevant to some of the challenges faced by instrumental variable methods.

## 1. INTRODUCTION

The task of performing graphical model selection arises in many applications in science and engineering. There are several factors that make this problem particularly challenging. First, it is common that only a subset of the relevant variables are observed and estimators that do not account for hidden variables are therefore prone to confounding. On the other hand, modelling latent variables is itself difficult because of identifiability and tractability issues. Second, the number of variables being modelled is often comparable to or greater than the number of samples. It is well-known that, in such a scaling regime, obtaining a consistent estimator is usually impossible without making further assumptions about the model, *e.g.* sparsity, low-dimensionality. Finally, modelling the joint distribution over all observed variables is not always relevant. It might be preferable to learn a graphical model over a subset of the collection while conditioning on the rest of the variables. This is the case in genetics, for example, where one could model a gene expression network conditional on the samples' DNA. Implicit to this idea is the desire to encode a notion of direction of effect: the expression of a gene can be influenced by DNA, but no changes in gene expression levels will modify an individual's genome. The result of such a discriminative approach is an increase in power and identifiability because explicit – and valid – assumptions are being made about the data.

An answer to some of these questions was given by (Chandrasekaran, Parrilo and Willsky 2012) who consider the problem of learning a graphical model in the presence of hidden variables. They suggest estimating an inverse covariance matrix which is the sum of a sparse and a low-rank matrix. It is this low-rank component which accounts for the marginal effect of the confounders. When the graphical model over the observed variables is sparse and there are few hidden variables with an effect spread over most nodes, they show that a consistent estimator for both matrices can be obtained and expressed as the solution to a convex problem. They further show that this approach behaves well in the high-dimensional setting. Unfortunately, the limitation on the number of hidden variables – for which the rank of the low-rank piece is a proxy – can be quite restrictive.

Another partial solution to our problem was introduced in (Sohn and Kim 2012) who defined the concept of a *sparse Gaussian conditional random field* (sparse GCRF), a regularised maximum likelihood estimator that learns a Gaussian graphical model over a subset of the variables ( $X$ , say) while conditioning on the remaining variables ( $Z$ , say). A sparse GCRF estimates two parameters

that correspond respectively to the direct actions of  $Z$  on  $X$  and to the structure of the graph over  $X$ . Multiple simulations and applications to computational biology and energy forecasting showed that such an approach performs substantially better than its counterparts in both model selection and prediction tasks, even in the high-dimensional case (Zhang and Kim 2014; Wytock and Kolter 2013). Of course, since hidden variables might induce incorrect edges, such results hold only if all the relevant variables are being observed

(Chandrasekaran, Parrilo and Willsky 2012) and (Sohn and Kim 2012) made significant progress on the challenges mentioned in the introductory paragraph but none of these methods can cope with the full problem. Here, we combine both approaches in order to learn a sparse GCRF in the presence of latent variables. Inputs (variables in  $Z$ ) are allowed to act on  $X$  in both a sparse and a low-rank fashion while the inverse covariance matrix over  $X$  is estimated conditional on  $Z$  and the latent variables. In this setting, latent factors can either mediate the effects of  $Z$  on  $X$  or simply act as confounders on  $X$ . By doing so we generalise the work done in (Chandrasekaran, Parrilo and Willsky 2012; Sohn and Kim 2012) but we also provide an alternative answer to some of the problems recently considered in the multivariate regression literature (Rothman, Levina and Zhu 2010), *e.g.* reduced rank-regression with inverse covariance estimation (Chen and Huang 2014). As will be shown later, this approach allows us to correctly recover graphs that are typically denser and with more hidden variables than the ones that can be handled by other methods.

From both a theoretical and a computational point of view, modelling latent variables in a discriminative fashion gives rise to a number of complications (*e.g.* proximal operator not defined in closed form) that we address in this paper. Under suitable conditions, however, this problem admits a unique solution that can be recovered as the minimum of a convex function. We derive convergence bounds for this estimator and show that it is well-behaved in the high-dimensional regime as well as “sparsistent” (*i.e.* capable of recovering the graph structure). We then describe a proximal gradient algorithm which is able to fit the model to thousands of variables. Through extensive simulations, we illustrate the conditions required for identifiability and show that there is a wide range of situations in which this model performs significantly better than its counterparts. For example, it can accommodate many more hidden variables. We also argue that our approach is relevant to tackle some of the problems faced by instrumental variables methods, *e.g.* coping with

invalid instruments (Kang, Zhang, Cai and Small n.d.; Bowden, Davey Smith and Burgess 2015).

## 2. PROBLEM STATEMENT

### 2.1 Setup

Throughout, we assume that we are given  $n$  independent, identically distributed, realisations of a Gaussian random vector  $Y \in \mathbb{R}^{m+p+h}$ .  $Y$  is indexed by disjoint subsets of  $\{1, \dots, m+p+h\}$ , denoted  $Z, X, H$  and with respective cardinality  $m, p$  and  $h$ . They correspond to the variables we wish to condition on, the variables we wish to model and the hidden variables. We write  $Y_Z$  (resp.  $Y_X$  and  $Y_H$ ) for the subvector of  $Y$  indexed by  $Z$  (resp.  $X$  and  $H$ ).  $\Sigma^*$  denotes the true covariance matrix over  $Z, X, H$ , and we define  $K^* = \Sigma^{*-1}$ , the true precision matrix. For any given matrix – such as  $K^* - K_Z^*, K_{ZX}^*, \dots$  represent the  $m \times m, m \times p, \dots$  submatrices extracted by considering only a subset of the rows/columns, as indicated by the subscripts. Furthermore,  $K_O^*$  represents the precision matrix indexed by all observed variables :  $O = Z \cup X$ . Finally, we write  $\Sigma^n$  for the sample covariance matrix :  $\frac{1}{n} \sum_i Y_i Y_i^T$ .

In this paper, we are concerned with learning a Gaussian graphical model (GGM). A graphical model is a model defined with respect to a graph whose nodes correspond to random variables and whose edges encode conditional independence statements among variables (Koller and Friedman 2009; Lauritzen 1996). A well-known property of GGMs (also known as Gaussian Markov Random Fields) is the connection existing between the inverse covariance matrix  $\Sigma^{*-1}$  (also known as precision or concentration matrix) and the structure of the (undirected) graphical model  $(V, E)$ : there is an edge between  $V_i$  and  $V_j$  if and only if  $\Sigma_{ij}^{*-1}$  is non-zero. This is a property on which many approaches rely.

### 2.2 Previous Work

As already pointed out, the relationship between concentration matrix and graph structure makes it possible to learn a sparse graph by controlling the number of non-zero entries of the estimate. In practice, model selection in the context of GGMs is often performed using an  $\ell_1$ -regularised maximum likelihood estimator (MLE) commonly known as the graphical lasso (Friedman, Hastie and Tibshirani 2008). The  $\ell_1$ -norm is the convex envelope of the  $\ell_0$  unit ball and is therefore a natural convex relaxation to learn sparse matrices. Building on the success of the graphical lasso,

estimators of the form “log-likelihood” + “non-Euclidian convex penalty” have received considerable interest (Chandrasekaran, Recht, Parrilo and Willsky 2012). A relevant example is the use of the nuclear norm (*i.e.* the sum of the singular values) as a convex relaxation for learning low-rank models (Bach 2008). Beyond their attractive computational properties, the  $\ell_1$  and nuclear norm regularised MLEs offer strong theoretical guarantees (Bach 2008; Ravikumar, Wainwright, Raskutti and Yu 2011). For example (Ravikumar et al. 2011) showed that, under well-defined circumstances, the graphical lasso is capable of correctly recovering a sparse graphical model.

The graphical lasso estimates a sparse precision matrix by modelling the joint distribution of a Gaussian vector. Taking this problem as a starting point, two elementary operations on multinormal distributions can be performed : conditioning and marginalising. The two approaches mentioned in the introduction arise as an answer to the challenges raised by each of these operations.

*Conditioning : Sparse Gaussian Conditional Markov Random Fields.* (Sohn and Kim 2012) consider the problem of learning a GGM (over  $X$ , say) conditional on a subset of the observed variables ( $Z$ , say). It is assumed that there are no hidden variables, so that  $H = \{\}$ . In this context, it is well-known that

$$Y_X|Y_Z \sim \mathcal{N}(-K_X^{*-1}K_{ZX}^{*T}Y_Z, K_X^{*-1}).$$

Consequently, (Sohn and Kim 2012) suggest the following estimator for  $K_{ZX}^*$  and  $K_X^*$ :

$$(\hat{S}_{ZX}, \hat{S}_X) = \arg \min_{S_X \succ 0, S_{ZX}} \ell(S_{ZX}, S_X; \Sigma_O^n) + \lambda_n(\|S_X\|_1 + \|S_{ZX}\|_1),$$

where

$$\ell(S_{ZX}, S_X; \Sigma_O^n) = -\log \det S_X + \text{Tr}(\Sigma_X^n S_X + 2\Sigma_{ZX}^n S_{ZX} + S_X^{-1} S_{ZX}^T \Sigma_Z^n S_{ZX}).$$

The entries of both  $S_X$  and  $S_{ZX}$  are being shrunk in order to jointly learn a pair of sparse matrices describing the direct effects of  $Z$  on  $X$  and the graph over  $X$ . (Wytock and Kolter 2013) studied the theoretical properties of this estimator and derived a set of sufficient conditions for the correct recovery of  $K_X^*$  and  $K_{ZX}^*$ . Through simulations and a number of applications, (Sohn and Kim 2012; Wytock and Kolter 2013; Zhang and Kim 2014) showed that, in some settings, this approach constitutes a significant improvement over the graphical lasso in both model selection and prediction tasks.

*Marginalising: Low Rank Plus Sparse Decomposition.* We now turn to the problem of learning a graphical model in the presence of hidden variables. We assume that only the variables indexed by  $X$  and  $Z$  are observed so that the marginal precision matrix is given by the Schur complement of  $H$  in  $K^*$ . Therefore,

$$Y_O \sim \mathcal{N} \left( 0, (K_O^* - K_{OH}^* K_H^{*-1} K_{OH}^{*T})^{-1} \right).$$

Here  $K_O^*$  would be our target : it encodes the structure of the graphical model over  $O$ , while  $K_{OH}^* K_H^{*-1} K_{OH}^{*T}$  encodes the marginal effect of all the hidden variables on  $O$ . Since the marginal precision matrix is the sum of two matrices ( $S - L$ , say), the problem is fundamentally misspecified. However, following the seminal work of (Candès, Li, Ma and Wright 2011; Chandrasekaran, Sanghavi, Parrilo and Willsky 2009), (Chandrasekaran, Parrilo and Willsky 2012) showed that it is sometimes possible to correctly decompose  $S - L$  into its summands and recover the structure of the graph encoded by  $S$ . Loosely speaking, this is the case if  $K_O^*$  is sparse and there are relatively few hidden variables with an effect spread over most of the nodes in  $O$ . As a result, (Chandrasekaran, Parrilo and Willsky 2012) suggest a regularised maximum likelihood estimator which penalises the  $\ell_1$ -norm of  $S$  and the nuclear norm of  $L$  as follows:

$$(\hat{S}, \hat{L}) = \arg \min_{S-L \succ 0, L \succeq 0} \ell(S - L; \Sigma^n) + \lambda_n(\gamma \|S\|_1 + \|L\|_*),$$

where  $\ell(K; \Sigma^n) = \text{Tr}(K \Sigma^n) - \log \det K$  and  $\lambda, \gamma > 0$ . With the help of the  $\ell_1$  and nuclear penalties, the precision matrix is therefore decomposed into the sum of a sparse and a low-rank matrix. Among other useful results, (Chandrasekaran, Parrilo and Willsky 2012) showed that this estimator is, under some suitable conditions, sparsistent and “ranksistent” : the sign patterns of both the entries of  $S$  and the spectrum of  $L$  can be recovered exactly. Modelling latent variables leads to two related, but distinct, problems:

- *identifiability* : when does the problem admit a *unique* solution? Notice that, unlike the breakdown caused by the high-dimensional regime, this kind of non-identifiability is more fundamental and remains no matter how large the number of samples.
- *consistency* : provided there exists a unique solution, when can the estimator correctly recover that solution?

In summary, in the approaches that we have described here, *conditioning* is modelled by using the relevant likelihood, while *marginalising* is handled by decomposing the marginal precision matrix into the sum of two components.

### 2.3 Suggested Estimator

In light of the work described above, we propose decomposing the parameters  $S_X$  and  $S_{ZX}$  of a Gaussian conditional Markov random field into the sum of a low-rank and a sparse matrix. To that end, we suggest optimising the following regularised likelihood

$$\begin{aligned}
(\hat{S}_X, \hat{L}_X, \hat{S}_{ZX}, \hat{L}_{ZX}) = & \arg \min_{S_X, L_X, S_{ZX}, L_{ZX}} \ell(S_{ZX} - L_{ZX}, S_X - L_X; \Sigma_O^n) + \lambda_n(\gamma \|S\|_1 + \|L\|_*) \\
& \text{s.t } S_X - L_X \succ 0, L_X \succeq 0 \text{ and } S = \begin{pmatrix} S_X \\ S_{ZX} \end{pmatrix}, L = \begin{pmatrix} L_X \\ L_{ZX} \end{pmatrix}, \quad (1)
\end{aligned}$$

where

$$\ell(K_{ZX}, K_X; \Sigma_O^n) = -\log \det K_X + \text{Tr}(\Sigma_X^n K_X + 2\Sigma_{ZX}^n K_{ZX} + K_X^{-1} K_{ZX}^T \Sigma_Z^n K_{ZX}).$$

Solving (1) amounts to solving a function which is *jointly convex* in its parameters over a *convex constraint set* (proofs are in Appendix B): an operation which can in general be performed in polynomial time. As mentioned earlier, this likelihood is structured around two parameters,  $K_{ZX}$  and  $K_X$ , accounting respectively for the direct (*i.e.* conditional on all variables) effects of  $Z$  on  $X$  and the structure of the graph over  $X$ . However, because we penalise the rank of  $L$ , the effect of all latent variables is modelled *jointly* and a single set of latent factors is learned. No distinction is being made between the variables that mediate the action of  $Z$  and the ones that act as confounders on  $X$ . On the other hand, the parameters  $S_X$  and  $S_{ZX}$  retain their interpretability.

The next two sections are dedicated to the study of the theoretical and computational properties of (1). At a theoretical level, our most important result is a proof of consistency for (1) showing the dependency of the estimator on the various properties of the problem, *e.g.* scaling regime, strength of the correlation between variables, etc... We then show how a proximal gradient algorithm can be used to find the solution to (1). Here, the main difficulty is the impossibility of computing a certain proximal operator in closed form.



### 3. THEORETICAL ANALYSIS

Earlier, we showed that the form taken by estimator (1) follows naturally from the study of i.i.d. realisations of a Gaussian vector, after conditioning on part of the variables and assuming that some variables are not observed. Later we will see that this is motivated by a number of real-life applications arising, for example, from the moralisation of some directed acyclic graphs.

For the time being, we simply assume that each sample is generated according to the model

$$Y_X|Y_Z \sim \mathcal{N}\left(-(S_X^* - L_X^*)^{-1}(S_{ZX}^* - L_{ZX}^*)^T Y_Z, (S_X^* - L_X^*)^{-1}\right), \quad (2)$$

and ask under what circumstances our estimator correctly recovers the parameters  $S^*, L^*$  (as built by stacking  $S_X^*, S_{ZX}^*$  and  $L_X^*, L_{ZX}^*$ ) with overwhelming probability.

We analyse this problem in the framework of (Chandrasekaran, Parrilo and Willsky 2012) and therefore our proofs often mirror theirs. However, because of the form taken by the likelihood and because we do not limit ourselves to positive semi-definite matrices, the analysis is substantially more complicated. Our goal being to provide the reader with an intuition for the range of applicability of our method, we choose to defer most technical details to the appendices and give only a few key results.

As mentioned earlier, modelling latent variables by decomposing the parameters into a sum of two matrices raises *identifiability* issues: given samples drawn from (2) when is it possible to exactly decompose the sum  $S - L$  (where  $S, L$  are defined as before) into its summands? It turns out that this is a problem which has been tackled in great generality in (Chandrasekaran, Parrilo and Willsky 2012) and their results apply directly to the present situation. We start by introducing a few notations and key definitions that will be used throughout the rest of this document. Then, an overview of the conditions required for identifiability is given and we look at our main result: the *consistency* of (1).

#### 3.1 Key Definitions and Assumptions

Until now, we have repeatedly mentioned the fact that the “low-rank plus sparse decomposition” was possible if  $S$  is sparse and  $L$  is low-rank. However, it is clear that imposing conditions on the sparsity of  $S$  and the rank of  $L$  is not sufficient. For example, consider the matrix with a single entry: it is at the same time sparse and low-rank and there is therefore no unique way of decomposing

it into the sum of a low-rank and a sparse matrix. (Chandrasekaran et al. 2009) introduce the notion of *rank-sparsity incoherence* and define quantities that make it possible to express clearly the conditions under which such a problem is well-posed. More details about the motivation behind these definitions are given in (Chandrasekaran, Parrilo and Willsky 2012).

**Key Definitions** We write  $\mathcal{S}(k)$  for the algebraic variety of  $(m+p) \times p$  matrices with at most  $k$  non-zero entries. Any  $(m+p) \times p$  matrix  $M$  with support of cardinality  $k$  is a smooth point of  $\mathcal{S}(k)$  and we denote by  $\Omega(M)$  the tangent space to  $\mathcal{S}(k)$  at  $M$  (this is the set of matrices in  $\mathcal{S}(k)$  whose support is included in the support of  $M$ ).

Similarly, we define  $\mathcal{L}(r)$ , the variety of  $(m+p) \times p$  matrices with rank at most  $r$ . Any matrix  $M$  of rank  $r$  is a smooth point of  $\mathcal{L}(r)$  and we write  $T(M)$  for the tangent space to  $\mathcal{L}(r)$  at  $M$  (this is the set of matrices in  $\mathcal{L}(r)$  whose row space is identical to  $M$ 's or whose column space is identical to  $M$ 's).

Given two linear spaces of identical dimension  $T_1, T_2$ , write

$$\rho(T_1, T_2) \triangleq \max_{\|N\|_2 \leq 1} \|(\mathcal{P}_{T_1} - \mathcal{P}_{T_2})(N)\|_2$$

for a measure of the “angle” between  $T_1$  and  $T_2$ . Here  $\mathcal{P}_{T_1}$  is the orthogonal projector onto  $T_1$  while  $\|N\|_2$  is the spectral norm of  $N$  (*i.e.* its largest singular value).  $\rho(T_1, T_2)$  is used to describe the set of tangent spaces that are close to the nominal  $T(L^*)$ .

(Chandrasekaran et al. 2009) quantify how spread the effect of the latent variables is on the observed variables by defining

$$\xi(T(M)) \triangleq \max_{N \in T(M), \|N\|_2 \leq 1} \|N\|_\infty$$

for any matrix  $M$ . Remark that  $\xi(T(M))$  controls all the elements of the tangent space at  $M$ . Thus, if  $\xi(T(M))$  is small, none of the matrices having the same row-space (resp. column-space) as  $M$  can have a large  $\ell_\infty$ -norm, meaning that the row-space (resp. column-space) of  $M$  cannot be closely aligned with any of the coordinate axes. Therefore, a small  $\xi(T(M))$  guarantees that no single latent variable will have a strong effect on only a small set of the observed variables (precisely because they define the coordinate axes).

Similarly, define

$$\mu(\Omega(M)) \triangleq \max_{N \in \Omega(M), \|N\|_\infty \leq 1} \|N\|_2,$$

to quantify how diffuse the spectrum of  $M$  is. One shows that matrices with a small number of non-zero entries per row/column (and thus sparse) have a small  $\mu(M)$ .

**Conditions on the Fisher Information Matrix** In the context of the graphical lasso, the *mutual incoherence* or *irrepresentability* condition is a well-known requirement for identifiability (Ravikumar et al. 2011). Just like the irrepresentability condition imposes restrictions on the Hessian of the likelihood, the conditions given in (Chandrasekaran, Parrilo and Willsky 2012) involve the Fisher Information Matrix (FIM)  $\mathcal{I}_{\Sigma_Z}^*$  evaluated at the true parameters  $S^* - L^*$ . The subscript is reminder that the entire analysis is performed conditional on  $\Sigma_Z$ . To avoid cluttered notations, we write  $\Omega = \Omega(S^*)$  and  $T = T(L^*)$  and denote by  $\mathcal{P}_\Omega, \mathcal{P}_T$  the orthogonal projections onto these linear subspaces.

We can now define the quantities that control the behaviour of the FIM when restricted to  $\Omega$  and  $T$ . Set

$$\begin{aligned} \alpha_\Omega &\triangleq \min_{M \in \Omega, \|M\|_\infty = 1} \|\mathcal{P}_\Omega \mathcal{I}_{\Sigma_Z}^* \mathcal{P}_\Omega(M)\|_\infty; \\ \delta_\Omega &\triangleq \max_{M \in \Omega, \|M\|_\infty = 1} \|\mathcal{P}_{\Omega^\perp} \mathcal{I}_{\Sigma_Z}^* \mathcal{P}_\Omega(M)\|_\infty, \\ \beta_\Omega &\triangleq \max_{M \in \Omega, \|M\|_2 = 1} \|\mathcal{I}_{\Sigma_Z}^*(M)\|_2. \end{aligned}$$

Likewise let,

$$\begin{aligned} \alpha_T &\triangleq \min_{\rho(T, T') < \xi(T)/2} \min_{M \in T', \|M\|_2 = 1} \|\mathcal{P}_{T'} \mathcal{I}_{\Sigma_Z}^* \mathcal{P}_{T'}(M)\|_2; \\ \delta_T &\triangleq \max_{\rho(T, T') < \xi(T)/2} \max_{M \in T', \|M\|_2 = 1} \|\mathcal{P}_{T'^\perp} \mathcal{I}_{\Sigma_Z}^* \mathcal{P}_{T'}(M)\|_2, \\ \beta_T &\triangleq \max_{\rho(T, T') < \xi(T)/2} \max_{M \in T', \|M\|_\infty = 1} \|\mathcal{I}_{\Sigma_Z}^*(M)\|_\infty. \end{aligned}$$

Finally, we define

$$\alpha \triangleq \min(\alpha_\Omega, \alpha_T) \quad \beta \triangleq \max(\beta_\Omega, \beta_T) \quad \delta \triangleq \max(\delta_\Omega, \delta_T).$$

Then the following assumption is a generalisation of the irrepresentability condition, and from now on, we assume that it holds.

**Assumption 1.** (*Generalised Irrepresentability Condition, (Chandrasekaran, Parrilo and Willsky 2012)*)

There exists a  $\nu \in (0, \frac{1}{2}]$  such that

$$\frac{\delta}{\alpha} \leq 1 - 2\nu.$$

A direct consequence of Assumption (1) is that  $\alpha > 0$ , so that the FIM is injective on  $\Omega$  and on all the spaces that are close to  $T$  (including  $T$  itself). For that reason, Assumption (1) subsumes the standard *restricted convexity assumption* which is also necessary in the context of sparse conditional Gaussian Markov random fields and regularised regression in order to guarantee that the likelihood – when restricted to the true support – is strictly convex, and therefore that there exists a unique optimum (Wytock and Kolter 2013; Wainwright 2009).

Another consequence of Assumption 1 is Proposition 1 given in Appendix A. This proposition is applied multiple times throughout the consistency proof. It holds only if the following conditions are met. This shows the role played by  $\mu$  and  $\xi$  in identifiability.

**Assumption 2.** (*Assumptions for Proposition 1*)

$$\mu(\Omega)\xi(T) \leq \frac{1}{6} \left( \frac{\nu\alpha}{\beta(2-\nu)} \right)^2$$

and  $\gamma$  is chosen in the range

$$\gamma \in \left[ \frac{3\xi(T)\beta(2-\nu)}{\nu\alpha}, \frac{\nu\alpha}{2\mu(\Omega)\beta(2-\nu)} \right].$$

Unsurprisingly, a sparse  $S^*$  (small  $\mu(\Omega)$ ) and a diffuse effect of the latent variables on the observed nodes (small  $\xi(T)$ ) increase the chances that Assumption (2) will hold.

### 3.2 Consistency

We can now present our main result and state the consistency of estimator (1) (see Appendix D for the proof).

We define the following quantities :  $D = \max(1, \frac{\nu\alpha}{3\beta(2-\nu)})$ ,  $\psi_Z = \|\Sigma_Z\|_2$ ,  $\psi_X^* = \|K_X^{*-1}\|_2$  and

$\phi_{ZX}^* = \|K_{ZX}^*\|_2$ ,  $\psi = \frac{3}{2}\psi_X^* \sqrt{\left(1 + 2\frac{\psi_Z}{\psi_X^*} (1 + \frac{9}{4}\psi_X^* \phi_{ZX}^*)^2\right)}$ . We also set,

$$\begin{aligned} C_1 &= \frac{48}{\alpha} + \frac{1}{\psi_X^{*2} \left(1 + 2\frac{\psi_Z}{\psi_X^*} (1 + \psi_X^* \phi_{ZX}^*)^2\right)}, \\ C_2 &= \left(\frac{24(2-\nu)}{\nu}\right) C_1^2 D \psi_X^{*2} \left(1 + 2\frac{\psi_Z}{\psi_X^*} (1 + \psi_X^* \phi_{ZX}^*)^2\right), \\ C_3 &= C_1 + \frac{3\alpha C_1^2 (2-\nu)}{4(3-\nu)}, C_4 = \max\{C_2, C_3\}, C_5 = \frac{C_1 \nu \alpha}{\beta(2-\nu)}, \\ C_6 &= \frac{\alpha \nu}{32(3-\nu)D} \min\left(\frac{1}{6\psi_X^*}, \frac{\phi_{ZX}^*}{4}, \frac{\alpha \nu}{384D(3-\nu)\psi_X^* \psi^2 (1 + \frac{\alpha}{6\beta})^2}\right). \end{aligned}$$

Finally, set:

$$\delta = 4\psi_X^* \max\left(4(1 + \psi_X^* \psi_Z \phi_{ZX}^{*2}), \psi_Z\right).$$

Then we prove the following theorem in the appendix:

**Theorem 1.** (*Algebraic Consistency*)

Suppose that Assumptions 1 and 2 hold and that we are given  $n$  samples drawn according to (2).

Further assume that the following hold:

1.  $n \geq \frac{mp}{\xi(T)^4} \max\left(2, \frac{256\psi_X^{*2}(1+\psi_X^* \psi_Z \phi_{ZX}^{*2})^2}{C_6^2}, \frac{16\psi_Z^2 \psi_X^{*2}}{C_6^2}\right)$ .
2. Set  $\lambda_n = \frac{6D\delta(2-\nu)}{\xi(T)\nu} \sqrt{\frac{mp}{n}}$ .
3. Let the minimum non-zero singular value of  $L^*$  be such that

$$\sigma \geq \frac{C_4 \lambda_n}{\xi(T)^2}.$$

4. Let the minimum magnitude nonzero entry  $\theta$  of  $S^*$  be such that

$$\theta \geq \frac{C_5 \lambda_n}{\mu(\Omega)}.$$

Then, with probability greater than  $1 - mp \exp(-mp)$  we have that

1.  $\text{sign}(\hat{S}) = \text{sign}(S^*)$  and  $\text{rank}(\hat{L}) = \text{rank}(L^*)$

2.

$$\max\left(\frac{1}{\gamma} \left\|\hat{S} - S^*\right\|_\infty, \left\|\hat{L} - L^*\right\|_2\right) \leq \frac{64(3-\nu)D}{\alpha \xi(T)\nu} \delta \sqrt{\frac{mp}{n}}.$$

Seen at a high-level, this result is analogous to the one obtained for the low-rank plus sparse decomposition of (Chandrasekaran, Parrilo and Willsky 2012) (*e.g.* scaling regime, dependency on  $\xi, \mu$ ), but our analysis also reveals important features of the problem that are specific to this setting.

Among others, there are some important consequences to this theorem:

1. This result holds even if  $m$  and  $p$  grow as a function of  $n$ .
2. Both the identifiability and consistency results depend on the structure of the graph and the number of hidden variables only through their dependence on  $\mu(\Omega)$  and  $\xi(T)$ .
3. Conditions c) and d) are fairly standard for this kind of estimator and but can be quite restrictive in practice. Through their dependency on  $\lambda_n$ , they vary with  $\frac{1}{\sqrt{n}}$ .
4. As expected, the predictive power of  $Z$  on  $X$  – as encoded by  $\phi_{ZX}^*$  – plays a key role in the convergence rate. More generally, the dependencies on the three important components of the model ( $Z$ , structure over  $X$ , relationship between  $Z$  and  $X$ ) is made clear by the dependency of  $\delta_n$  on  $\psi_Z, \psi_X^*$  and  $\phi_{ZX}^*$ .

We remark that because we model all latent variables jointly (with  $L$ ) our analysis does not really distinguish between  $S_X, L_X, L_{ZX}$  and  $S_{ZX}$ . Specifically, our proof makes a recurrent use of the identity

$$\max(\|A\|_2, \|B\|_2) \leq \left\| \begin{pmatrix} A \\ B \end{pmatrix} \right\|_2 \leq \sqrt{2} \max(\|A\|_2, \|B\|_2),$$

hence the form taken by  $\delta_n$ . Whether it may be possible to decompose the contribution of these terms is not yet clear.

#### 4. OPTIMISATION

Here we regard (1) as an optimisation problem and suggest a strategy to fit the model when  $m$  and  $p$  are in the thousands. As we show below, this is a fairly standard problem which has been extensively studied in the literature. This is why we do not give details about the full procedure. Instead, we focus on a difficulty which is specific to (1): the lack of closed-form solution for a certain proximal operator.

Seen at a high level, (1) is a well-studied problem of the form

$$\text{minimize } f(x) + g(x) \quad (3)$$

where  $f : \mathbb{R}^k \rightarrow \mathbb{R}$ ,  $g : \mathbb{R}^k \rightarrow \bar{\mathbb{R}}$  are closed proper convex functions and  $f$  is differentiable.

In this context, we have

$$f : (K_{ZX}, K_X) \mapsto -\log \det K_X + \text{Tr} \left( \Sigma_X^n S_X + 2\Sigma_{ZX}^n K_{ZX} + K_X^{-1} K_{ZX}^T \Sigma_Z^n K_{ZX} \right),$$

and

$$g : (K_{ZX}, K_X) \mapsto \lambda(\gamma \|S\|_1 + \|L\|_*) + \mathcal{I}_{\mathcal{S}^{\succ 0}}(S - L) + \mathcal{I}_{\mathcal{S}^{\succeq 0}}(L),$$

where  $S = \begin{pmatrix} S_X \\ S_{ZX} \end{pmatrix}$  and  $L = \begin{pmatrix} L_X \\ L_{ZX} \end{pmatrix}$ . Here we defined  $\mathcal{I}_{\mathcal{S}^{\succ 0}}$  as the indicator function of the cone

$$\mathcal{S}^{\succ 0} \triangleq \left\{ \begin{pmatrix} K_X \\ K_{ZX} \end{pmatrix} \mid K_X \in \mathbb{R}^{p \times p}, K_{ZX} \in \mathbb{R}^{m \times p}, K_X \succ 0 \right\}.$$

$\mathcal{I}_{\mathcal{S}^{\succeq 0}}$  is defined similarly for positive semi-definite matrices.

The non-differentiability of  $g$  prevents the resort to standard optimisation methods, which is why (3) is typically solved using some form of proximal algorithm (Parikh and Boyd 2014). The simplest instance of such an algorithm is a proximal gradient method with iteration

$$x^{k+1} := \text{prox}_{\eta^k g}(x^k - \eta^k \nabla f(x^k)), \quad (4)$$

where  $\eta^k$  is some step size and where  $\text{prox}_{\eta g}$  is the *proximal operator* of  $\eta g$ :

$$\text{prox}_{\eta g}(v) \triangleq \arg \min_x \eta g(x) + \frac{1}{2} \|x - v\|_2^2.$$

This is a well-known algorithm which has been extended and improved in many ways (*e.g.* Nesterov's method which was used to solve a similar problem in (Zhang and Kim 2014; Nesterov 2005)). However, no matter which flavour of the algorithm is being used, computing the proximal operator fast and accurately is critical. In particular, (Schmidt, Roux and Bach 2011) show how convergence is affected when an error is present in the calculation of the proximal operator and provide guidelines about the number of inner iterations required to achieve the optimal convergence rate.

This is relevant to our setting because one shows that applying iteration (4) to solve (1) requires the computation of a proximal operator of the form

$$\text{prox}_{\mathcal{I}_{\mathcal{S} \geq 0}(\cdot) + \lambda \|\cdot\|_*}(X) = \arg \min_L \mathcal{I}_{\mathcal{S} \geq 0}(L) + \lambda \|L\|_* + \frac{1}{2} \|L - X\|_F^2, \quad (5)$$

which is the proximal operator of the sum of two functions : one is the indicator function of a convex set, while the other corresponds to the regularisation imposed on the nuclear norm of  $L$ . Considered independently, the proximal operators of  $\mathcal{I}_{\mathcal{S} \geq 0}(L)$  and  $\lambda \|L\|_*$  admit well-known closed form solutions that rely on the thresholding of the eigenvalues of  $L_X$  and the singular values of  $L$ , respectively (Parikh and Boyd 2014). The proximal operator of their sum, however, cannot be computed in closed form. We must resort to the “Dykstra-like” proximal algorithm of (Combettes and Pesquet 2009) which generates a sequence that converges to the solution of (5). See Algorithm 1 for more details.

Initialisation: Set  $X_0 = X$ ,  $P_0 = 0$ ,  $Q_0 = 0$ ;

**for**  $k = 1, \dots, K$  **do**

$Y_k = \text{prox}_{\mathcal{I}_{\mathcal{S} \geq 0}}(X_k + P_k);$
$P_{k+1} = X_k + P_k - Q_k;$
$X_{k+1} = \text{prox}_{\lambda \ \cdot\ _*}(Y_k + Q_k);$
$Q_{k+1} = Y_k + Q_k - X_{k+1}$

**end**

**Result:**  $X_k$ .

**Algorithm 1:** “Dykstra-like” algorithm of (Combettes and Pesquet 2009). The sequence  $X_k$  converges to the solution of problem (5). In practice,  $\frac{\|X_{k+1} - X_k\|_F}{\|X_k\|_F} \leq \epsilon$  is used as a stopping criterion. We choose a fixed  $\epsilon$  of  $10^{-5}$ , but an adaptive scheme might result in faster convergence rates (Schmidt et al. 2011). Typically, each iteration has time complexity  $O(\max(p, m)^3)$ .

As a result, each iteration of (4) requires the computation of the gradient of  $f$  (with time complexity  $O(p^3)$ ), the proximal operator of the  $\ell_1$ -norm which is a simple thresholding of the entries and a few iterations of Algorithm 1 – typically less than 10.

In practice, we use a Nesterov’s second method – which is a form of accelerated proximal gradient algorithm – that we complement with the adaptive restart method introduced in (O’Donoghue and



Candès 2013; Nesterov 2005). In spite of the difficulties raised by the form of (5) this approach provably converges to the optimum of (1) (Schmidt et al. 2011). Another option would be to use the Alternating Direction Method of Multipliers (ADMM) but this would also involve some form of Algorithm 1 (Boyd 2011).

## 5. SIMULATIONS

We study the properties of the proposed model on synthetic data and compare its performances to the three other methods introduced earlier: the graphical lasso (Friedman et al. 2008), the sparse conditional Gaussian random Markov field (Zhang and Kim 2014; Sohn and Kim 2012; Wytock and Kolter 2013) and the low-rank plus sparse decomposition (Chandrasekaran, Parrilo and Willsky 2012). Each of these methods is subject to identifiability issues which can be caused either by a high-dimensional breakdown or because multiple models can describe the data equivalently well (*i.e.* are Markov equivalent). Here, our focus is on the latter issue: we draw a large number of samples from a range of models and measure their ability to recover the structure of the underlying graph.

### 5.1 Graphical Structures

For all simulations, each observation is generated according to a model of the form

$$Y_X|Y_{ZH} \sim \mathcal{N}(-S_X^{*-1}S_{ZX}^{*T}Y_{ZH}, S_X^{*-1}).$$

To account for the sparse/low-rank behaviour of the inputs and the hidden variables we generate  $Y_Z$  and  $Y_H$  according to various models that we define below.

The structure of the graph over  $X$ ,  $S_X^*$  is identical for all simulations: it consists of a chain of length  $p$  in which one link out of five has been removed :  $S_{X,ij}^* = S_{X,ji}^* = \rho$  if  $j = i + 1$  and  $i \not\equiv 0 \pmod{5}$ . Also,  $S_{X,ii}^* = 1$ . Figure 1 shows what the true precision matrix looks like, along with the corresponding covariance matrix.

$S_{ZX}^*$  is always constructed as follows. It is a matrix of size  $2p \times p$  such that  $S_{ZX}^* = [\beta\rho A_p \ \gamma\rho B_p]^T$ , where  $A_p, B_p$  are *random* matrices obtained by permuting the columns of  $I_p$ .  $\beta$  and  $\gamma$  control the strength of the effect of  $Z$  and  $H$ , respectively. The permutations are here to guarantee that the inputs and hidden variables are connected to a random subset of  $X$ .

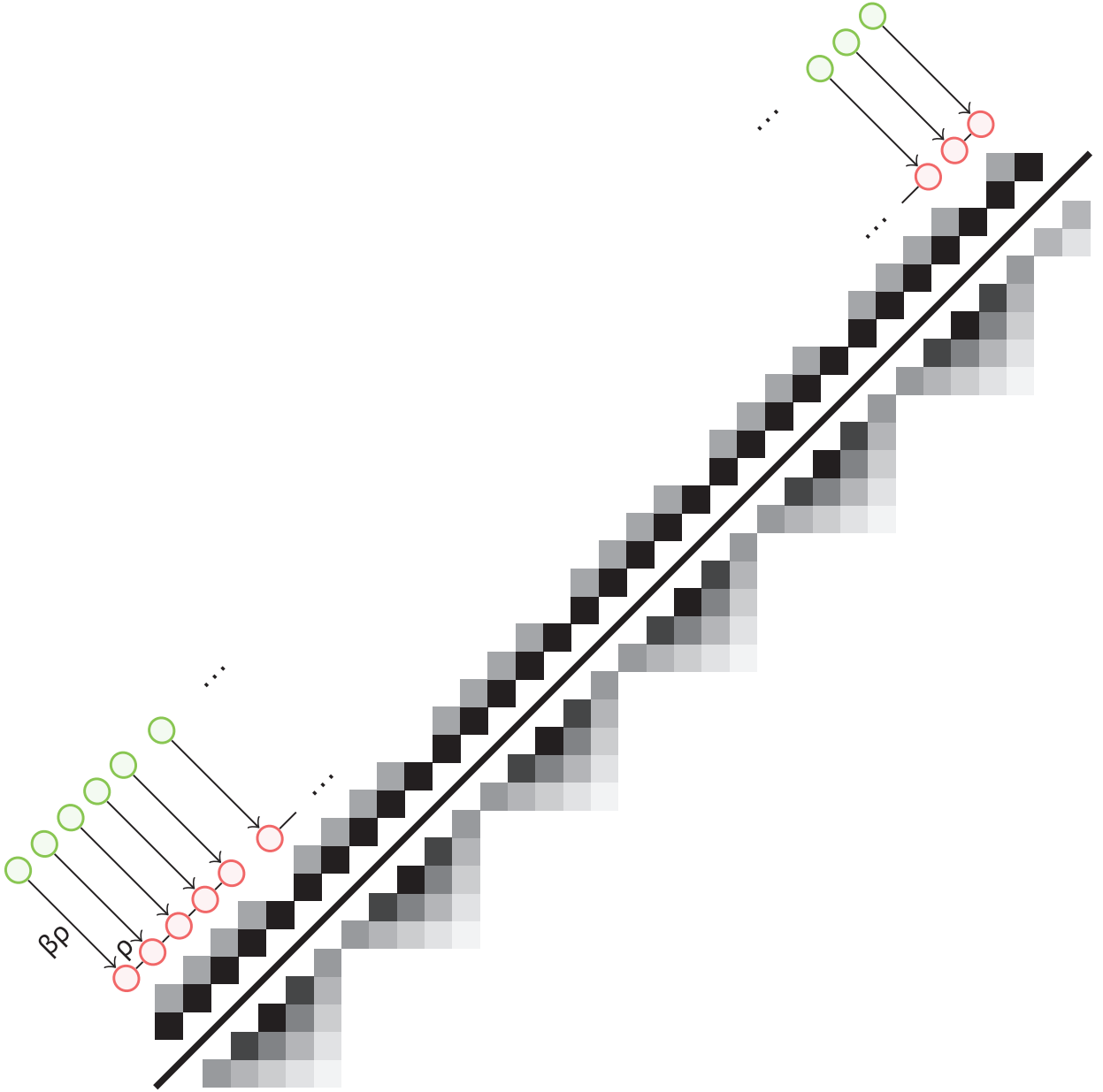


Figure 1:  $\mathcal{G}_{55}$  : Conditional graphical model in which the inputs (in green) are in a one-to-one mapping with the outputs (in red). There are no latent variables. The upper triangular matrix (above the solid black line) depicts the *precision* matrix  $S_X$  that we seek to recover. It corresponds to the red nodes of  $\mathcal{G}_{55}$ . The lower triangular matrix is the corresponding *covariance* matrix.

To account for the sparsity/low-rank structure of  $Z$  and  $H$ , we generate  $Y_Z$  and  $Y_H$  according to a simple tree model. We assume that  $p$  is of the form  $p = 2^k$ , pick an integer  $d_Z \in \{0, \dots, k\}$  and let  $Y_Z^{(2^{d_Z})} \in \mathbb{R}^{2^{d_Z}}$  be a standard Gaussian random vector (*i.e.* the covariance matrix is the identity). Then,  $Y_Z^{(2^{d_Z+1})}$  is defined inductively according to

$$Y_Z^{(2^{d_Z+1})} = \begin{pmatrix} Y_Z^{(2^{d_Z})} \\ Y_Z^{(2^{d_Z})} \end{pmatrix}.$$

We set  $Y_Z = Y_Z^{(2^k)}$  and assume that only  $Y_Z^{(2^{d_Z})}$  is observed.  $Y_H$  is defined according to the same process (replace  $d_Z$  with  $d_H$ ).

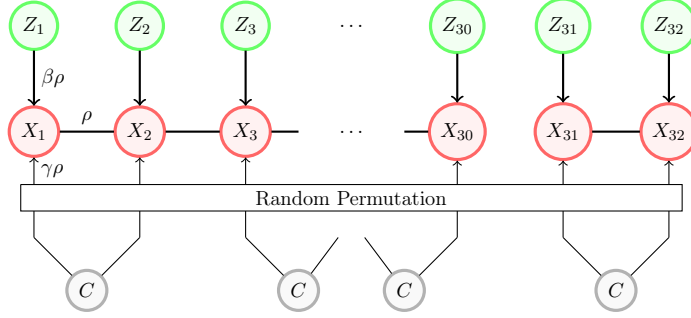
If  $d_Z = 0$  then a single input has an effect on all the variables in  $X$ . On the other hand, if  $d_Z = k$ , then there are  $k$  distinct inputs that are in a one-to-one correspondence with the variables in  $Z$ . If  $d_Z = k - 1$ , then each input acts on two random nodes of  $X$ , thus the “effective”  $S_{ZX}$  is fairly sparse. By varying  $d_Z$  and  $d_H$ , we control the sparsity/low-dimensionality of the inputs and the latent variables. For example, whenever,  $d_H < k$ , hidden variables confound  $S_X$ .

From now on, we write  $\mathcal{G}_{d_Z, d_H}$  for the graphical structure generated by choosing some pair  $(d_Z, d_H)$ . Figure 2 shows the structure of the graphs for different values of  $d_Z, d_H$ . As  $d_Z$  and  $d_H$  depart from  $k$ , the relationship between  $Z$  and  $X$  – along with the structure over  $X$  – become harder to estimate.

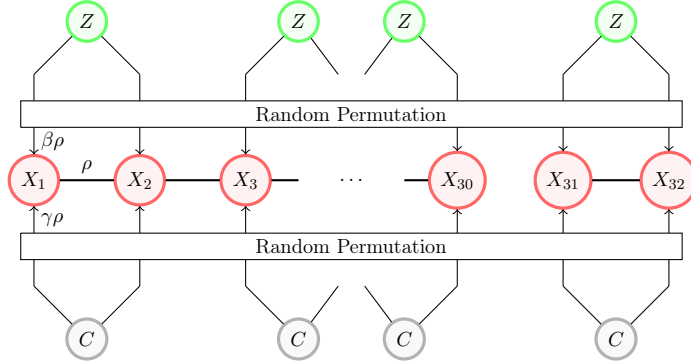
## 5.2 Results

In our simulations, we set  $p = 2^5, \beta = \gamma = 5, \rho = 0.2$  and  $n = 5 \cdot 10^5$ . We then generate data according to  $\mathcal{G}_{d_Z, d_H}$ , letting  $d_Z$  and  $d_H$  range from 5 to 2. At one end of the spectrum ( $d_Z = d_H = 5$ ) there is no confounding and the relationship between  $Z$  and  $X$  is easy to learn, while at the other end ( $d_H = 2$ ) there are only a few confounders with an effect spread over many observed variables: this falls directly within the range of applicability of the low-rank plus sparse decomposition. The intermediate regime is where our method is more likely to be beneficial: the effect of  $Z$  is not so sparse, confounding is not so incoherent. By generating data from these 16 models we try to provide a minimal set of designs that span the range of situations one might face in real applications.

The performance of all four methods (graphical lasso, sparse conditional GGM, low-rank plus sparse, our method) across the 16 simulation designs is shown in Figure 3. Here, we are interested



(a)  $\mathcal{G}_{54}$ , graph corresponding to  $d_Z = 5, d_H = 4$ . Each confounder (grey nodes) acts on two random observed variables (in red). With  $p = 32$ , there are 16 confounders.



(b)  $\mathcal{G}_{44}$ , graph corresponding to  $d_Z = 4, d_H = 4$ . Inputs (green nodes) act on two random observed variables.

Figure 2: Graphical representation of the models described by  $\mathcal{G}_{54}$  and  $\mathcal{G}_{44}$ , for  $p = 32$ . Other structures are defined inductively. For example  $\mathcal{G}_{32}$  would be the graph with  $2^3$  inputs (each acting on 4 outputs) and  $2^2$  confounders (each acting on 8 random outputs).

in recovering the structure of  $S_X$  and use precision / recall curves as a metric. A precision and a recall of 1 indicate consistency. The LR+S and LSCGGM both have two tuning parameters ( $\lambda$  and  $\gamma$ ). For each value of  $\gamma$ , one obtains a different precision/recall curve by varying  $\lambda$ . Figure 3 shows the curve obtained by selecting  $\gamma$  in order to maximise recall at a precision of 1. Only results for  $\rho = 0.2$  are shown here but similar experiments have been conducted for  $\rho = 0.05, 0.1$  and gave similar results.

First, we see that known methods behave as expected. The graphical lasso achieves consistency when there is no confounding and  $Z$  acts in a sparse fashion. The sparse conditional GGM approach is more robust to changes in  $S_{ZX}$  but this is restricted to situations in which there is no confounding. The low-rank plus sparse (LR+S) method is consistent when there are few latent variables (four, for a total of 32 observed variables); this corresponds to  $d_H = 2$ . Whenever any of these three methods is consistent, our method (LSCGGM) offers comparable performances.  $d_H = 4$  corresponds to the extreme situation in which each latent variable confounds exactly two random variables.

More importantly, we identify a regime ( $d_H = 3$ ) in which LSCGGM outperforms the other methods. In this regime, there are 8 latent variables (each latent variable affects 4 random observed nodes), so that the effect of the hidden variables is not so incoherent. Such a regime is far less restrictive than the one required for the consistency of the LRPS method, thus showing that we benefit from the additional assumptions made about the data.

As mentioned above, both the LR+S method and the suggested approach (LSCGGM) have two tuning parameters:  $\lambda$  and  $\gamma$ . While  $\lambda$  controls the overall shrinkage on the sparsity/rank of the estimates,  $\gamma$  accounts for the trade-off between sparse and low-rank components. In order to understand the effect of  $\gamma$  on the regularisation paths of these methods, we study the surface formed by the set of precision/recall curves obtained for various values of  $\gamma$ . Following the suggestion made in (Chandrasekaran et al. 2009) we reparametrise the penalty term as  $\lambda(\gamma\|S\|_1 + (1 - \gamma)\|L\|_*)$ , so that  $\gamma$  ranges from 0 to 1 instead of  $(0, +\infty)$ . Figure 4 illustrates how precision/recall curves evolve as one changes the relative weight given to the  $\ell_1$ /nuclear norm. By analogy to the Area Under Curve (AUC) metric, we measure the “Volume Under Surface” which accounts for the effect of both regularisation parameters.

Firstly, Figure 4 provides an illustration of some of the theoretical results of the previous section:

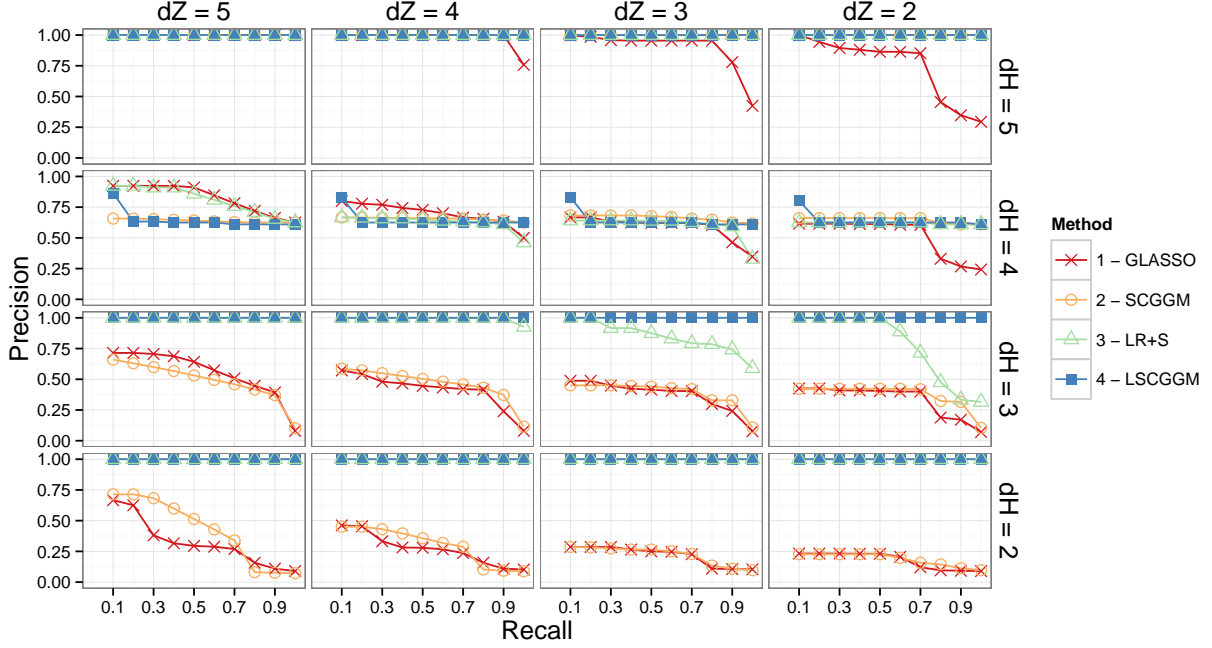


Figure 3: Comparison of the suggested estimator (LSCGGM) to other published methods. Along the  $x$ -axis (resp  $y$ -axis),  $d_Z$  (resp  $d_H$ ) varies from 5 to 2. More precisely, in the first row, there is no confounding at all. In the second row, hidden variables act in a very sparse fashion, and none of the methods are able to recover  $S_X$ . In the last row, there are 4 hidden variables and we are in the range of applicability of the low-rank plus sparse method. The third row corresponds to an intermediate regime in which there are twice as many latent variables. The suggested method outperforms the LR+S approach and achieves consistency in a wider range of situations. Settings:  $p = 2^5$ ,  $n = 5.10^5$ ,  $\beta = \gamma = 5$ ,  $\rho = 0.2$ . Each design is repeated 25 times. The value of the tuning parameter  $\gamma$  was chosen to maximise recall at a precision of 1 (see Figure 4 for the full surface in the cases  $d_Z = 3$ ,  $d_H = 2, 3$ ). Similar experiments were conducted with other values of  $\rho$  and gave similar results.

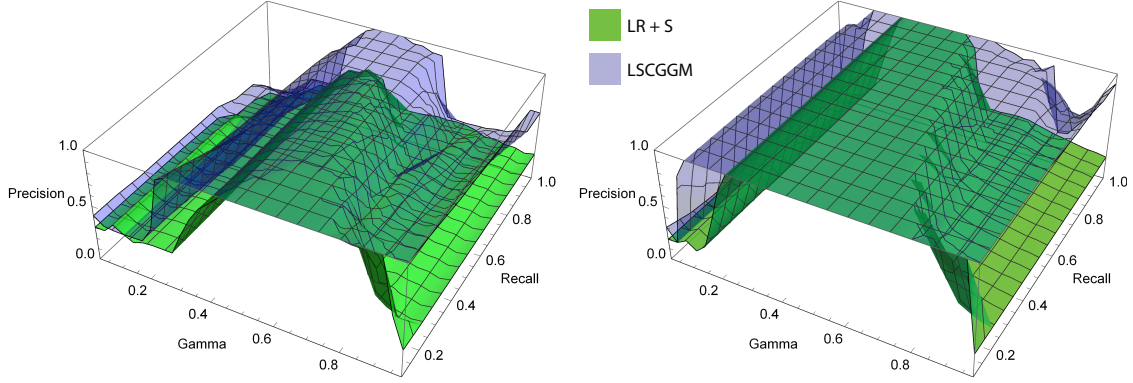
when achieved, consistency holds for a range of values of  $\gamma$ . Furthermore, for almost all designs, this range is much greater for the LSCGGM method than for the LR+S approach, thus reducing the sensitivity to this tuning parameter. This makes LSCGGM easier to use in practice and offers greater guarantees at all recall values.

## 6. DISCUSSION

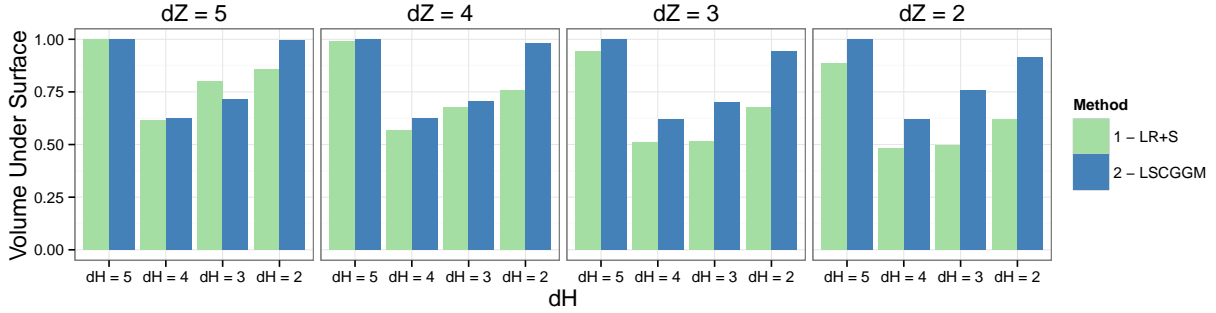
In this paper, we discussed the problem of estimating a conditional Gaussian graphical model in the presence of latent variables. Building on the framework previously introduced by the authors of (Chandrasekaran, Parrilo and Willsky 2012), we suggested an estimator – expressed as the minimum of a convex function – which decomposes the parameters of a conditional Gaussian graphical model into the sum of a sparse and a low-rank matrix. The proposed approach comports a single low-rank matrix and learns all latent variables jointly – irrespective of whether they act as confounders or mediate the effect of the variables being conditioned on.

Our main theoretical result is a proof of consistency for estimator (1). We showed that under some circumstances it is possible to recover the structure of a sparse conditional graphical model with overwhelming probability and with bounded errors terms. Furthermore, this result holds in the high-dimensional regime, *i.e.* when the the number of variables (whether being modelled or conditioned on) grows as a function of the sample size. On the computational side, we showed how off-the-shelf proximal gradient algorithms can be used to fit the model when the number of variables is in the thousands. Through simulations we provided an intuition about the various situations in which our estimator, and competing approaches, achieve consistency.

A relevant application is the one of learning a causal network in the presence of many “putative” instrumental variables. Instrumental variables have been used in fields such as econometrics and genetic epidemiology in order to test (and estimate) causal relationships between pairs of observed variables. Roughly speaking, one is typically interested in knowing whether  $X_1$  has a causal effect on  $X_2$  using  $Z_1, \dots, Z_m$  as instruments. For the variables  $Z = (Z_i)_{i \in \{1, \dots, m\}}$  to be *valid* instruments the distribution  $P(X_1, X_2, Z, U)$  ( $U$  describes the latent variables that might confound the study)



(a) Precision/recall surface for  $\mathcal{G}_{23}$  (*i.e.* each input  $Z$  acts on 8 random outputs and there are confounding variables). Both methods are consistent. As predicted from theory, this is achieved for a range of values of  $\gamma$ . The suggested method is less sensitive to this tuning parameter.

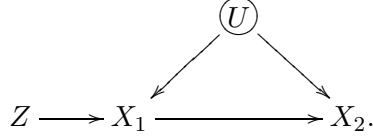


(c) Volume under surface across all 16 simulation designs. In almost of situations, the suggested approach is less sensitive to the value of  $\gamma$ .

Figure 4: Sensitivity to the tuning parameter  $\gamma$ . Here, an alternative parametrisation of the regularisation term is used:  $\lambda_n(\gamma\|S\|_1 + (1 - \gamma)\|L\|_*)$ . As a result,  $\gamma$  is in the range  $(0, 1)$  instead of  $(0, +\infty)$ : as  $\gamma \rightarrow 0$ , the penalty on the sparse component vanishes and one obtains a denser graph with no latent variables. Similarly, letting  $\gamma \rightarrow 1$  yields a very sparse graph with many latent variables. To each value of  $\gamma$  corresponds a precision/recall curve obtained by varying  $\lambda$ . At the top, the surface plots illustrate the “precision/recall surfaces” (for 30 values of  $\gamma \in [0.02, 0.98]$ ) in two special cases :  $\mathcal{G}_{23}$  and  $\mathcal{G}_{22}$ . Bottom: comparison of the volume under the precision/recall surfaces in all 16 simulation designs.



has to factor over the following graph:



Establishing that  $P(X_1, X_2, Z, U)$  actually factorises over this graph is challenging because it is not testable (see (Kang et al. n.d.; Bowden et al. 2015) for novel results on this problem). Now, if there are many observed variables  $X_1, \dots, X_p$  generated according to some linear causal model, a related problem is the one learning the moralised graph of that causal model, conditioning on  $Z$ . Using our approach, we decompose the parameters as:

$$\begin{pmatrix} S_X \\ S_{ZX} \end{pmatrix} - \begin{pmatrix} L_X \\ L_{ZX} \end{pmatrix}.$$

Because of moralisation,  $L_{ZX}$  is non-zero. However, one shows that the rank of  $L$  is equal to  $L_X$  so that adding instruments (provided they act sparsely on  $X$ ) increases the chances of the problem being identified. Thus, using  $\hat{S}_{ZX}$  and pairwise independence tests, it is possible to distinguish the valid instruments from the invalid ones.

Naturally, the method suggested here also suffers from a number of limitations and more work is required. For example, assuming that the latent variables are normally distributed appears quite restrictive when compared to the flexibility offered by instrumental variable methods. The question of learning discrete graphical models is also important but it is not yet clear how the present work can be extended to such models.

## A. IDENTIFIABILITY

As already mentioned in the theoretical section of the paper, the question of identifiability has already been answered in (Chandrasekaran, Parrilo and Willsky 2012). The reason for this is that only a number of properties of the Fisher Information Matrix appear in the proofs but the actual form of the FIM is irrelevant.

Here our goal is to recall a Proposition of (Chandrasekaran, Parrilo and Willsky 2012) which is instrumental in the consistency proof and to which we will refer time and again.

We first introduce a few additional notations. First, let  $g_\gamma$  denote the dual norm of the regularisation function  $f_\gamma = \lambda(\gamma \|S\|_1 + \|L\|_*)$ :

$$g_\gamma = \max \left( \frac{\|S\|_\infty}{\gamma}, \|L\|_2 \right).$$

We also write  $\mathcal{A} : \mathbb{R}^{(m+p) \times p} \times \mathbb{R}^{(m+p) \times p} \rightarrow \mathbb{R}^{(m+p) \times p}$  for the operator that adds two matrices. As usual,  $\mathcal{A}^\dagger$  denotes its adjoint :  $\mathcal{A}^\dagger : M \mapsto (M, M)$ , so that  $\mathcal{A}^\dagger \mathcal{A}(A, B) = (A + B, A + B)$ .

Then, under Assumptions 1 and 2 we have the following result.

**Proposition 1.** (*(Chandrasekaran, Parrilo and Willsky 2012), Proposition 3.3*)

Let  $\mathcal{Y} = \Omega \times T'$  with  $\rho(T, T') \leq \frac{\xi(T)}{2}$ . Then

1.

$$\min_{(S, L) \in \mathcal{Y}, \|S\|_\infty = \gamma, \|L\|_2 = 1} g_\gamma(\mathcal{P}_\mathcal{Y} \mathcal{A}^\dagger \mathcal{I}_{\Sigma_Z}^* \mathcal{A} \mathcal{P}_\mathcal{Y}(S, L)) \geq \frac{\alpha}{2}$$

and specifically, for  $(S, L) \in \mathcal{Y}$

$$g_\gamma(\mathcal{P}_\mathcal{Y} \mathcal{A}^\dagger \mathcal{I}_{\Sigma_Z}^* \mathcal{A} \mathcal{P}_\mathcal{Y}(S, L)) \geq \frac{\alpha}{2} g_\gamma(S, L).$$

2. Writing  $\mathcal{Y}^\perp$  for the orthogonal complement of  $\mathcal{Y}$ , we have

$$\left\| \mathcal{P}_{\mathcal{Y}^\perp} \mathcal{A}^\dagger \mathcal{I}_{\Sigma_Z}^* \mathcal{A} \mathcal{P}_\mathcal{Y} (\mathcal{P}_\mathcal{Y} \mathcal{A}^\dagger \mathcal{I}_{\Sigma_Z}^* \mathcal{A} \mathcal{P}_\mathcal{Y})^{-1} \right\|_{g_\gamma \rightarrow g_\gamma} \leq 1 - \nu.$$

And, more specifically,

$$g_\gamma(\mathcal{P}_{\mathcal{Y}^\perp} \mathcal{A}^\dagger \mathcal{I}_{\Sigma_Z}^* \mathcal{A} \mathcal{P}_\mathcal{Y}(S, L)) \leq (1 - \nu) g_\gamma(\mathcal{P}_\mathcal{Y} \mathcal{A}^\dagger \mathcal{I}_{\Sigma_Z}^* \mathcal{A} \mathcal{P}_\mathcal{Y}(S, L)).$$

## B. ELEMENTARY PROPERTIES OF THE LIKELIHOOD

We now make a number of straightforward calculations about the likelihood

$$\ell(K_{ZX}, K_X; \Sigma_O^n) = -\log \det K_X + \text{Tr}(\Sigma_X^n K_X + 2\Sigma_{ZX}^n K_{ZX} + K_X^{-1} K_{ZX}^T \Sigma_Z^n K_{ZX}).$$

We are interested in computing the gradient and the Hessian of  $\ell$ . In passing, we show that  $\ell$  is *convex*, so that optimisation problem (1), being the sum of three convex functions, is also convex.

**Property 1.** (*Maximum Likelihood Estimate (M.L.E.)*)

Assuming  $\Sigma_Z^*$  is non-singular and  $n$  is large enough, the M.L.E. is well-defined and given by

$$\begin{aligned}\hat{K}_X &= (\Sigma_X - \Sigma_{ZX}^T \Sigma_Z^{-1} \Sigma_{ZX})^{-1}; \\ \hat{K}_{ZX} &= -\Sigma_Z^{-1} \Sigma_{ZX} \hat{K}_X.\end{aligned}$$

We can now compute the Hessian of  $\ell$ .

**Property 2.** (*Fisher Information Matrix*)

$$\mathcal{I}_{\Sigma_Z}^* = - \begin{pmatrix} K_X^{*-1} \otimes K_X^{*-1} & 0 \\ 0 & 0 \end{pmatrix} - 2K_X^{*-1} \otimes \begin{pmatrix} K_X^{*-1} K_{ZX}^{*T} \Sigma_Z^* K_{ZX}^* K_X^{*-1} & -K_X^{*-1} K_{ZX}^{*T} \Sigma_Z^* \\ \cdot & \Sigma_Z \end{pmatrix}.$$

*Proof.* We use differentials.

First, it is easy to see that

$$\begin{aligned}d\ell &= -\text{tr}(K_X^{-1} dK_X) + \text{tr}(\Sigma_X dK_X) + 2\text{tr}(\Sigma_{ZX} (dK_{ZX})^T) \\ &\quad + 2\text{tr}(\Sigma_Z dK_{ZX} K_X^{-1} K_{ZX}^T) - \text{tr}(\Sigma_Z K_{ZX} K_X^{-1} dK_X K_X^{-1} K_{ZX}^T).\end{aligned}$$

Likewise, it is straightforward to compute the second order derivatives. To make the computation easier to follow, we break down  $d^2\ell$  into its individual components:

$$\begin{aligned}d(-\text{tr}(K_X^{-1} dK_X)) &= \text{tr}(K_X^{-1} dK_X A_X^{-1} dK_X); \\ d(\text{tr}(\Sigma_X dK_X)) &= 0; \\ d(\text{tr}(\Sigma_{ZX} (dK_{ZX})^T)) &= 0; \\ d(\text{tr}(\Sigma_Z dK_{ZX} K_X^{-1} K_{ZX}^T)) &= -\text{tr}(K_X^{-1} dK_X K_X^{-1} K_{ZX}^T \Sigma_Z dK_{ZX}) \\ &\quad + \text{tr}(K_X^{-1} dK_{ZX}^T \Sigma_Z dK_{ZX}); \\ d(-\text{tr}(\Sigma_Z K_{ZX} K_X^{-1} dK_X K_X^{-1} K_{ZX}^T)) &= 2\text{tr}(K_X^{-1} dK_X K_X^{-1} K_{ZX}^T \Sigma_Z K_{ZX} K_X^{-1} dK_X) \\ &\quad - 2\text{tr}(K_X^{-1} dK_{ZX}^T \Sigma_Z K_{ZX} K_X^{-1} dK_X).\end{aligned}$$

So that:

$$\begin{aligned} d^2\ell &= \text{tr}(K_X^{-1}dK_XK_X^{-1}dK_X) + 2\text{tr}(K_X^{-1}dK_{ZX}^T\Sigma_ZdK_{ZX}) \\ &\quad + 2\text{tr}(K_X^{-1}dK_XK_X^{-1}K_{ZX}^T\Sigma_ZK_{ZX}K_X^{-1}dK_X) \\ &\quad - 4\text{tr}(K_X^{-1}dK_{ZX}^T\Sigma_ZK_{ZX}K_X^{-1}dK_X). \end{aligned}$$

Now, write  $K := \begin{pmatrix} K_X \\ K_{ZX} \end{pmatrix}$  and similarly  $\text{vec}(K) := \begin{pmatrix} \text{vec}(K_X) \\ \text{vec}(K_{ZX}) \end{pmatrix}$ . Then, using the identity  $\text{tr}(ABCD) = (\text{vec}B^T)^T(A^T \otimes C)\text{vec}D$ , we obtain

$$\begin{aligned} \text{tr}(K_X^{-1}dK_XK_X^{-1}dK_X) &= d(\text{vec}K_X)^T(K_X^{-1} \otimes K_X^{-1})d\text{vec}K_X \\ \text{tr}(K_X^{-1}dK_{ZX}^T\Sigma_ZdK_{ZX}) &= d(\text{vec}K_{ZX})^T(K_X^{-1} \otimes \Sigma_Z)d\text{vec}K_{ZX} \\ \text{tr}(K_X^{-1}dK_XK_X^{-1}K_{ZX}^T\Sigma_ZK_{ZX}K_X^{-1}dK_X) &= d(\text{vec}K_X)^T(K_X^{-1} \otimes K_X^{-1}K_{ZX}^T\Sigma_ZK_{ZX}K_X^{-1})d\text{vec}K_X \\ \text{tr}(K_X^{-1}dK_{ZX}^T\Sigma_ZK_{ZX}K_X^{-1}dK_X) &= d(\text{vec}K_{ZX})^T(K_X^{-1} \otimes \Sigma_ZK_{ZX}K_X^{-1})d\text{vec}K_X. \end{aligned}$$

So that,

$$\begin{aligned} d^2\ell &= d(\text{vec}K_X)^T [(K_X^{-1} \otimes K_X^{-1}) + 2(K_X^{-1} \otimes K_X^{-1}K_{ZX}^T\Sigma_ZK_{ZX}K_X^{-1})] d\text{vec}K_X \\ &\quad + 2d(\text{vec}K_{ZX})^T(K_X^{-1} \otimes \Sigma_Z)d\text{vec}K_{ZX} \\ &\quad - 4d(\text{vec}K_{ZX})^T(K_X^{-1} \otimes \Sigma_ZK_{ZX}K_X^{-1})d\text{vec}K_X \end{aligned}$$

and

$$d^2\ell = (d\text{vec}K)^T \begin{pmatrix} K_X^{-1} \otimes K_X^{-1} + 2K_X^{-1} \otimes K_X^{-1}K_{ZX}^T\Sigma_ZK_{ZX}K_X^{-1} & -2K_X^{-1} \otimes \Sigma_ZK_{ZX}K_X^{-1} \\ & 2K_X^{-1} \otimes \Sigma_Z \end{pmatrix} d\text{vec}K.$$

From which the result follows.  $\square$

In passing, remark that  $\mathcal{I}_{\Sigma_Z}(K)$  is positive semi-definite if and only if

$$\begin{pmatrix} K_X^{-1}K_{ZX}^T\Sigma_ZK_{ZX}K_X^{-1} & -K_X^{-1}K_{ZX}^T\Sigma_Z \\ & \Sigma_Z \end{pmatrix}$$

is positive semi-definite. This is because  $K_X$  is positive definite (we have the constraint  $K_X \succ 0$ .)

But it's easy to see that this matrix is simply a covariance matrix of the form  $A^T A$ . In conclusion,

$\ell$  is convex and so is (1).

## C. CURVATURE OF THE LIKELIHOOD AND THE RANK VARIETY

### C.1 Curvature of the rank variety

The goal of this section is to extend two results of (Chandrasekaran, Parrilo and Willsky 2012; Chandrasekaran 2011) from the special case of positive definite matrices to arbitrary matrices. We rely heavily on the previous work of (Bach 2008).

We give here the results that we seek to prove and dedicate the rest of this section to their proof.

Given two linear subspaces  $T1, T2$  of same dimension we define

$$\rho(T1, T2) \triangleq \|\mathcal{P}_{T1} - \mathcal{P}_{T2}\|_{2 \rightarrow 2} = \max_{\|N\|_2 \leq 1} \|(\mathcal{P}_{T1} - \mathcal{P}_{T2})(N)\|_2,$$

which measures the "angle" between these two subspaces. Recall that for any matrix  $M \in \mathbb{R}^{p \times q}$ ,  $T(M)$  denotes the tangent space to the variety of low-rank matrices at  $M$ . Then we have the following proposition

**Proposition 1.** (*Extension of (Chandrasekaran 2011), Proposition 4.2.1, Proposition 4.2.2*)

Let  $W \in \mathbb{R}^{p \times q}$  be a rank  $r < \min(p, q)$  matrix with non-zero singular values  $(\sigma_i)_{i \in \{1, \dots, r\}}$ . Write  $\sigma = \min_i \sigma_i$ , and let  $\Delta$  be such that  $\|\Delta\|_2 < \frac{\sigma}{4}$ . Let  $W + \Delta$  be rank  $r$  matrix. Then we have

1.

$$\rho(T(W + \Delta), T(W)) \leq \frac{8}{\sigma} \|\Delta\|_2;$$

2.

$$\left\| \mathcal{P}_{T(W)^\perp}(\Delta) \right\|_2 \leq \frac{4}{\sigma} \|\Delta\|_2^2.$$

**Jordan-Wieland Matrix and Matrix perturbation bounds** Throughout, we assume that  $W$  is some  $\mathbb{R}^{p \times q}$  matrix, with non-zero singular values  $\sigma_i, i = 1, \dots, r$  (indexed by decreasing order) and singular vectors,  $u_i, v_i$ . The corresponding Jordan-Wiedlandt matrix is defined from  $W$  as follows (Stewart and Sun 1990):

$$\bar{W} \begin{pmatrix} 0 & W \\ W^T & 0 \end{pmatrix}.$$

This matrix is interesting because it has eigenvalues  $\sigma_i, -\sigma_i$  and its eigenvectors are  $\frac{1}{\sqrt{2}} \begin{pmatrix} u_i \\ \pm v_i \end{pmatrix}$ .

We write  $\bar{W} = \bar{U} \bar{S} \bar{U}^T$  for the eigenvalue decomposition of  $\bar{W}$ . If the singular value decomposition

of  $W$  is  $W = USV^T$ , then we have  $\bar{S} = \frac{1}{\sqrt{2}} \begin{pmatrix} S & 0 \\ 0 & -S \end{pmatrix}$  and  $\bar{S} = \frac{1}{\sqrt{2}} \begin{pmatrix} U & U \\ V & -V \end{pmatrix}$  so that

$$\bar{U}\bar{U}^T = \begin{pmatrix} UU^T & 0 \\ 0 & VV^T \end{pmatrix}.$$

Now, let  $\Delta$  be a small perturbation to  $W$ . If  $\|\Delta\|_2 \leq \frac{\sigma_r}{2}$  then  $W + \Delta$  has  $r$  singular values that are strictly greater than  $\sigma_r/2$  and all the remaining ones are strictly less than  $\sigma_r/2$  (Stewart and Sun 1990). Now, let  $P_{\bar{W}, \sigma_r/2}$  denote the projector on the  $p + q - 2r$ -dimensional invariant subspace of  $\bar{W}$  which corresponds to the smallest eigenvalues (in this case it is a null space as  $W$  is of rank exactly  $r$ ). Likewise, if  $\|\Delta\|_2 \leq \sigma_r/2$ ,  $P_{\bar{W} + \bar{\Delta}, \sigma_r/2}$  also denotes a projector onto a  $p + q - 2r$  dimensional subspace. The projection onto the orthogonal subspace is given by  $I - P_{\bar{W} + \bar{\Delta}, \sigma_r/2}$ .

(Bach 2008) shows the following results

**Proposition 2.** ((Bach 2008), Proposition 16)

Assume  $W$  is of rank  $k$  and  $\|\Delta\|_2 < \frac{\sigma_r}{4}$ . Then, the projection on the first  $r$  eigenvectors of  $\bar{W}$ ,  $I - P_{\bar{W}, \sigma_r/2}$ , is such that

$$\left\| P_{\bar{W} + \bar{\Delta}, \sigma_r/2} - P_{\bar{W}, \sigma_r/2} \right\|_2 \leq \frac{4}{\sigma_r} \|\Delta\|_2.$$

Now, recall that for any two matrices  $A, B$  (of compatible dimensions) we have the following inequalities:

$$\max(\|A\|_2, \|B\|_2) \leq \|C\|_2 \leq \|A\|_2 + \|B\|_2,$$

where  $C = \begin{pmatrix} A \\ B \end{pmatrix}$ . Writing  $P_{U(W)}$  (resp.  $P_{V(W)}$ ) for the projection onto the row-space  $U(W)$  (resp.  $V(W)$ )) then we have the following corollary

**Corollary 1.** Assume  $W$  is of rank  $k$  and  $\|\Delta\|_2 < \frac{\sigma_r}{4}$ . Assume further that  $\text{rank}(W + \Delta) = r$ . Then we have that

$$\max\left(\|P_{U(W+\Delta)} - P_{U(W)}\|_2, \|P_{V(W+\Delta)} - P_{V(W)}\|_2\right) \leq \left\| P_{\bar{W} + \bar{\Delta}, \sigma_r/2} - P_{\bar{W}, \sigma_r/2} \right\|_2.$$

We also have the following result:

**Proposition 3.** ((Bach 2008), Proposition 18)

Assume that  $W$  has rank  $r < \min(p, q)$ , with singular value decomposition  $W = USV^T$ . If  $\frac{4}{\sigma_r} \|\Delta\|_2^2 < \|(I - UU^T)\Delta(I - VV^T)\|_2$ , then  $\text{rank}(W + \Delta) > r$ .

Curvature of the rank variety – Proof of Proposition 1 Let  $W$  be a  $p \times q$  matrix defined as before. The projection onto the matrix variety of low-rank matrices at  $W$ ,  $T(W)$ , is written  $\mathcal{P}_{T(W)}$ . For any matrix  $N$ , it can be expressed using the row/column projectors as follows ((Chandrasekaran et al. 2009)):

$$\mathcal{P}_{T(W)}(N) = P_{U(W)}N + NP_{V(W)} - P_{U(W)}NP_{V(W)},$$

while the projector onto the orthogonal subspace is  $(I - \mathcal{P}_{T(W)})$ , *i.e.*:

$$\mathcal{P}_{T(W)^\perp} = (I - P_{U(W)}N(I - P_{V(W)}).$$

For any matrix  $N$  we have that

$$\begin{aligned} (\mathcal{P}_{T(W+\Delta)} - \mathcal{P}_{T(W)}) = \\ (P_{U(W+\Delta)} - P_{U(W)}) N (I - P_{V(W)}) + (I - P_{U(W+\Delta)}) N (P_{V(W+\Delta)} - P_{V(W)}). \end{aligned}$$

As a result, under the assumptions of Proposition 1 (in particular that  $\|\Delta\|_2 \leq \sigma_r/4$ ), we have the following inequalities:

$$\begin{aligned} \rho(T(W + \Delta), T(W)) &\leq \max_{\|N\|_2 \leq 1} \| (P_{U(W+\Delta)} - P_{U(W)}) N (I - P_{V(W)}) + \\ &\quad (I - P_{U(W+\Delta)}) N (P_{V(W+\Delta)} - P_{V(W)}) \|_2 \\ &\leq \max_{\|N\|_2 \leq 1} \| (P_{U(W+\Delta)} - P_{U(W)}) N (I - P_{V(W)}) \|_2 + \\ &\quad \max_{\|N\|_2 \leq 1} \| (I - P_{U(W+\Delta)}) N (P_{V(W+\Delta)} - P_{V(W)}) \|_2 \\ &\leq 2 \max(\|P_{U(W+\Delta)} - P_{U(W)}\|, \|P_{V(W+\Delta)} - P_{V(W)}\|) \\ &\leq \frac{8}{\sigma_r} \|\Delta\|_2, \end{aligned}$$

where we used the corollary given earlier. This proves part 1) of Proposition 1.

We now turn to part 2). This is a direct consequence of Proposition 2. Indeed, we have that

$$\|\mathcal{P}_{T(W)^\perp}(\Delta)\|_2 = \|(I - P_{U(W)}\Delta(I - P_{V(W)})\|_2,$$

which – using the contra-positive of Proposition 2 – concludes the proof.

Curvature of the likelihood Throughout the rest of the appendix we will make use of the following notations. They have all been defined in the main paper as well as in (Chandrasekaran, Parrilo and Willsky 2012).

Let  $w = \max(1, \frac{1}{\gamma})$  and set  $D = \max(1, \frac{\nu\alpha}{3\beta(2-\nu)})$ . We assume that

$$\gamma \in \left[ \frac{3\xi(T)\beta(2-\nu)}{\nu\alpha}, \frac{\nu\alpha}{2\mu(\Omega)\beta(2-\nu)} \right],$$

so that  $w \leq \frac{D}{\xi(T)}$ .

In order to bound the error terms, it is necessary to study the curvature of the likelihood gradient at the true parameters. In particular, we want to know how “close” (here taken in the  $\|\cdot\|_2$  sense) to the true parameters the estimates have to be in order to achieve a particular error bound.

To that end we introduce the following matrix valued function which maps, for a given input covariance matrix, the parameters to the gradient of the likelihood function.

$$\begin{aligned} \mathcal{F}_{\Sigma_Z} : \quad \mathbb{R}^{(m+p) \times p} &\rightarrow \mathbb{R}^{(m+p) \times p} \\ M = \begin{pmatrix} M_X \\ M_{ZX} \end{pmatrix} &\mapsto \begin{pmatrix} \mathcal{F}_{\Sigma_Z, X}(M) \\ \mathcal{F}_{\Sigma_Z, ZX}(M) \end{pmatrix} \triangleq \begin{pmatrix} M_X^{-1} + M_X^{-1} M_{ZX}^T \Sigma_Z M_{ZX} M_X^{-1} \\ -2\Sigma_Z M_{ZX} M_X^{-1} \end{pmatrix}. \end{aligned}$$

We also define the following matrix valued function which maps the sample covariance matrix to the expression of the MLE (provided it is defined). :

$$\begin{aligned} \mathcal{H} : \mathcal{S}_{(m+p)}^{>0} &\rightarrow \mathcal{S}_{(m+p)}^{>0} \\ \Sigma &\mapsto \begin{pmatrix} (\Sigma_X - \Sigma_{ZX}^T \Sigma_Z^{-1} \Sigma_{ZX})^{-1} \\ -\Sigma_Z^{-1} \Sigma_{ZX} (\Sigma_X - \Sigma_{ZX}^T \Sigma_Z^{-1} \Sigma_{ZX})^{-1} \end{pmatrix}. \end{aligned}$$

In particular, we have

$$\mathcal{F}_{\Sigma_Z} \mathcal{H} \Sigma^* = \begin{pmatrix} \Sigma_X^* - \Sigma_{ZX}^{*T} \Sigma_Z^{*-1} (\Sigma_Z^{-1} \Sigma_Z^{*-1} - I) \Sigma_{ZX}^* \\ 2\Sigma_Z^{-1} \Sigma_Z^{*-1} \Sigma_{ZX}^* \end{pmatrix}.$$

We can now express the following proposition which will be used in later in order to bound the error terms.

**Proposition 4.** *Consider the Taylor expansion of  $\mathcal{F}_{\Sigma_Z}$  at  $K^* \triangleq \mathcal{H} \Sigma^*$ :*

$$\mathcal{F}_{\Sigma_Z}(K^* + \Delta) = \mathcal{F}_{\Sigma_Z} K^* + \mathcal{I}_{\Sigma_Z}^* \Delta + R_{K^*}(\Delta),$$

(note that in order to avoid cluttered notations we do not write  $R_{K^*, \Sigma_Z}(\Delta)$ , but the reader should not forget that there is a dependency here.). Write  $\psi_Z \triangleq \|\Sigma_Z\|_2$ ,  $\psi_X^* \triangleq \|K_X^{*-1}\|_2$  and  $\phi_{ZX}^* \triangleq \|K_{ZX}^*\|_2$ .

If  $\|\Delta\|_2 \leq \min(\frac{1}{3\psi_X^*}, \frac{\phi_{ZX}^*}{2})$  then

$$\|R_{K^*}(\Delta)\|_2 \leq 3\psi_X^* \psi^2 \|\Delta\|_2^2,$$



where

$$\psi \triangleq \frac{3}{2}\psi_X^* \sqrt{\left(1 + 2\frac{\psi_Z}{\psi_X^*} \left(1 + \frac{9}{4}\psi_X^* \phi_{ZX}^*\right)^2\right)}.$$

*Proof.* We start by recalling that for any two matrices  $A, B$  (of compatible dimensions) we have the following inequalities:

$$\max(\|A\|_2, \|B\|_2) \leq \|C\|_2 \leq \|A\|_2 + \|B\|_2,$$

where  $C = \begin{pmatrix} A \\ B \end{pmatrix}$ .

By the mean value theorem, there exists a real  $t$ ,  $0 \leq t \leq 1$ , such that  $R_{K^*}(\Delta) = d^2(\mathcal{F}_{\Sigma_Z}(K^* + t\Delta); \Delta)$ , which is the second derivative at  $K^* + t\Delta$  evaluated at  $\Delta$ . Given the inequality above, it is clear that, in order to obtain a bound on the norm of the remainder, it is enough to bound the sum  $\|R_{X,K^*}(\Delta_X)\|_2 + \|R_{ZX,K^*}(\Delta_{ZX})\|_2$ . This is achieved by bounding both terms individually.

A tedious calculation (similar in all points to the one given in the appendix of (Wytock and Kolter 2013)) yields the following expressions:

$$\begin{aligned} d^2(\mathcal{F}_{X,\Sigma_Z}(M); \Delta) &= 2(M_X^{-1}\Delta_X M_X^{-1}\Delta_X M_X^{-1} + M_X^{-1}\Delta M_X^{-1}\Delta M_X^{-1} M_{ZX}^T \Sigma_Z^* M_{ZX} M_X^{-1} \\ &\quad M_X^{-1}\Delta_X M_X^{-1} M_{ZX}^T \Sigma_Z^* M_{ZX} M_X^{-1} \Delta_X M_X^{-1} + \\ &\quad M_X^{-1} M_{ZX}^T \Sigma_Z^* M_{ZX} M_X^{-1} \Delta_X M_X^{-1} \Delta_X M_X^{-1} - \\ &\quad M_X^{-1} \Delta_X M_X^{-1} \Delta_{ZX}^T \Sigma_Z^* M_{ZX} M_X^{-1} - M_X^{-1} \Delta_X M_X^{-1} M_{ZX}^T \Sigma_Z^{-1} \Delta_{ZX} M_X^{-1} - \\ &\quad M_X^{-1} \Delta_{ZX} \Sigma_Z^* M_{ZX} \Delta_X M_X^{-1} - M_X^{-1} M_{ZX}^T \Sigma_Z^* \Delta_{ZX} M_X^{-1} \Delta_X M_X^{-1} + \\ &\quad M_X^{-1} \Delta_{ZX}^T \Sigma_Z^* \Delta_{ZX} M_X^{-1}). \end{aligned}$$

$$d^2(\mathcal{F}_{ZX,\Sigma_Z}(M); \Delta) = 2(-2\Sigma_Z^* M_{ZX} M_X^{-1} \Delta_X M_X^{-1} \Delta_X M_X^{-1} + 2\Sigma_Z^* \Delta_{ZX} M_X^{-1} \Delta_X M_X^{-1})$$

We are interested in bounding  $\|d^2(\mathcal{F}_{\Sigma_Z}(K^* + t\Delta); \Delta)\|_2$  and the terms  $\|(K_X^* + t\Delta_X)^{-1}\|_2$ ,  $\|K_{ZX}^* + t\Delta_{ZX}\|_2$  appear many times in these expressions. We use the assumption on  $\|\Delta\|_2$  to show that  $R_{K^*,\Sigma_Z}(\Delta)$  converges and to bound these two terms.

- Rewrite  $(K_X^* + t\Delta_X)^{-1}$  as  $K_X^{*-1}(I + tK_X^{*-1}\Delta_X)^{-1}$ . Using the submultiplicative property of the spectral norm and the fact that  $(I + tK_X^{*-1}\Delta_X)^{-1} = \sum_{i=0}^{\infty} (-1)^i (tK_X^{*-1}\Delta_X)^i$ , we obtain :

$$\|(K_X^* + t\Delta_X)^{-1}\|_2 \leq \psi_X^* \frac{1}{1 - \psi_X^* \|\Delta_X\|_2};$$

which, by our assumptions on  $\|\Delta\|_2$ , implies

$$\|(K_X^* + t\Delta_X)^{-1}\|_2 \leq \frac{3}{2}\psi_X^*.$$

- On the other hand, we have

$$\|K_{ZX}^* + t\Delta_{ZX}\|_2 \leq \|K_{ZX}^*\|_2 + \|\Delta_{ZX}\|_2 \leq \frac{3}{2}\phi_{ZX}^*.$$

Using these two inequalities, it is now straightforward to bound the remainder by bounding its summands independently. Putting together similar terms and rewriting the expression, we have

$$\|R_{K^*}(\Delta)\|_2 \leq 2\left(\frac{3}{2}\psi_X^*\right)^3 \left(1 + 3\frac{\psi_Z}{\frac{3}{2}\psi_X^*} \left(1 + \left(\frac{3}{2}\right)^2 \psi_X^* \phi_{ZX}^*\right)^2\right) \max(\|\Delta_X\|_2^2, \|\Delta_{ZX}\|_2^2),$$

which completes the proof.  $\square$

As a corollary, we can prove a result which is similar to Proposition B.4.1 in (Chandrasekaran 2011).

**Corollary 2.** *Suppose that  $\gamma$  is in the range required for identifiability. Let  $g_\gamma(\Delta_S, \Delta_L) \leq \min(\frac{1}{3\psi_X^*}, \frac{\phi_{ZX}^*}{2})\frac{1}{1+\frac{\alpha}{6\beta}}$ , for any  $(\Delta_S, \Delta_L)$  with  $\Delta_S \in \Omega$ . Then we have*

$$g_\gamma(\mathcal{A}^\dagger R_{K^*} \mathcal{A}(\Delta_S, \Delta_L)) \leq 3\frac{D}{\xi(T)}\psi_X^*\psi^2(1 + \frac{\alpha}{6\beta})^2 g_\gamma(\Delta_S, \Delta_L)^2.$$

*Proof.* Following the proof given in (Chandrasekaran, Parrilo and Willsky 2012; Chandrasekaran 2011), we derive a result that relates the dual norm  $g_\gamma$  to the spectral norm :

$$\begin{aligned} \|\mathcal{A}(\Delta_S, \Delta_L)\|_2 &\leq \|\Delta_S\|_2 + \|\Delta_L\|_2 \\ &\leq \gamma\mu(\Omega)\frac{\|\Delta_S\|_\infty}{\gamma} + \|\Delta_L\|_2 \\ &\leq (1 + \gamma\mu(\Omega))g_\gamma(\Delta_S, \Delta_L) \\ &\leq (1 + \frac{\alpha}{6\beta})g_\gamma(\Delta_S, \Delta_L) \\ &\leq \min(\frac{1}{3\psi_X^*}, \frac{\phi_{ZX}^*}{2}), \end{aligned}$$

where we used the range of  $\gamma$  and the assumption on  $g_\gamma$ . Therefore, the assumptions of Proposition 4 are met and we have:

$$\|R_{K^*}(\mathcal{A}(\Delta_S, \Delta_L))\|_2 \leq 3\psi_X^*\psi^2\|\mathcal{A}(\Delta_S, \Delta_L)\|_2^2$$

which implies

$$\|R_{K^*}(\mathcal{A}(\Delta_S, \Delta_L))\|_2 \leq 3\psi_X^*\psi^2(1 + \frac{\alpha}{6\beta})^2 g_\gamma(\Delta_S, \Delta_L)^2,$$

from which the result follows.  $\square$

#### D. CONSISTENCY

This section studies a variant of estimator of (1) in which the constraint  $L_X \succeq 0$  is removed:

$$\begin{aligned} (\hat{S}_X, \hat{L}_X, \hat{S}_{ZX}, \hat{L}_{ZX}) = \\ \arg \min_{S_X, L_X, S_{ZX}, L_{ZX}} \ell(S_{ZX} - L_{ZX}, S_X - L_X; \Sigma_O^n) + \lambda(\gamma \|S\|_1 + \|L\|_*) \\ \text{s.t } S_X - L_X \succ 0 \text{ and } S = \begin{pmatrix} S_X \\ S_{ZX} \end{pmatrix}, L = \begin{pmatrix} L_X \\ L_{ZX} \end{pmatrix}. \end{aligned} \quad (\text{A.1})$$

Our proof of consistency follows the same pattern as the one in (Chandrasekaran, Parrilo and Willsky 2012). Because the likelihood is different and because we consider rectangular matrices (the upper part of which are positive definite), the actual statements of the theorems and propositions are almost always different. In some cases, however, the proof is easily adapted from (Chandrasekaran, Parrilo and Willsky 2012) without any major difficulty. For that reason, we will focus on points that are critical to this particular analysis.

Compared to the standard approach to prove the consistency of lasso-type estimators, such as the one used in (Ravikumar et al. 2011; Wytock and Kolter 2013), this proof strategy is more involved. Indeed, since the variety of sparse matrices has zero-curvature in its smooth points, tangent space constraints and variety constraints are equivalent. Unfortunately, we also have to deal with the rank variety which has non-zero curvature, hence the need to control for all the tangent spaces that are close to the nominal  $T$ .

A detailed description of the proof strategy, along with the rationale behind this approach, is given in (Chandrasekaran, Parrilo and Willsky 2012). The key steps are the following:

- We start by considering a version of (A.1) in which the tangent space constraints are enforced explicitly. In its resort Brouwer's fixed point theorem, this part is fairly similar to the proof given in (Ravikumar et al. 2011; Wytock and Kolter 2013).

- We then consider a problem in which the variety constraints are explicitly enforced and show that the optimum of that non-convex problem is algebraically consistent.
- We then show that the optima of the variety constrained and tangent space constrained problems are identical.
- Finally, we derive conditions for the optimum of (A.1) to be identical to the optimum of the tangent space constrained problem.

#### D.1 Tangent space constraints

We consider

$$\begin{aligned}
(\hat{S}_\Omega, \hat{L}_{T'}) &= \arg \min_{S, L} -\log \det(S_X - L_X) + \text{tr}(\Sigma_X(S_X - L_X)) + 2\text{tr}(\Sigma_{ZX}(S_{ZX} + L_{ZX})^T) \\
&\quad + \text{tr}(\Sigma_Z(S_{ZX} + L_{ZX})(S_X - L_X)^{-1}(S_{ZX} + L_{ZX})^T) \\
&\quad \text{s.t. } S_X - L_X \succ 0, S \in \Omega, L \in T' \\
&\quad \text{where } S = \begin{pmatrix} S_X \\ S_{ZX} \end{pmatrix}, L = \begin{pmatrix} L_X \\ L_{ZX} \end{pmatrix};
\end{aligned} \tag{A.2}$$

for some  $T'$ . The goal of this section is to show that if  $T'$  is sufficiently close to the nominal  $T$  (as measured by  $\rho(T, T')$ ), then the error terms are bounded.

Let  $C_{T'} = \mathcal{P}_{T'}(L^*)$  be the orthogonal projection of the nominal low-rank matrix onto the linear subspace  $T'$ .

**Proposition 5.** *Let the errors  $(\Delta_S, \Delta_L)$  be defined as above and assume that  $T'$  is such that  $\rho(T, T') \leq \frac{\xi(T)}{2}$ . Define*

$$E_n = \begin{pmatrix} \Sigma_X^n \\ 2\Sigma_{ZX}^n \end{pmatrix} - \mathcal{F}_{\Sigma_Z^n}(\mathcal{H}\Sigma^*)$$

and

$$r = \max \left( \frac{8}{\alpha} \left( g_\gamma(\mathcal{A}^\dagger E_n) + g_\gamma(\mathcal{A}^\dagger \mathcal{I}_K^* C_{T'}) + \lambda_n \right), \|C_{T'}\|_2 \right).$$

If

$$r \leq \min \left( \frac{1}{6\psi_X^*}, \frac{\phi_{ZX}^*}{4}, \frac{\alpha\xi(T)}{96D\psi_X^*\psi^2(1 + \frac{\alpha}{6\beta})^2} \right),$$

then

$$g_\gamma(\Delta_S, \Delta_L) \leq 2r.$$

*Proof.* The proof is very similar to the one given in (Chandrasekaran, Parrilo and Willsky 2012).

The only difficulty is to rewrite the difference

$$\begin{pmatrix} \Sigma_X^n \\ 2\Sigma_{ZX}^n \end{pmatrix} - \mathcal{F}_{\Sigma_Z^n}(\hat{S}_\Omega - \hat{L}_{T'})$$

in terms of the errors  $(\Delta_S, \Delta_L)$ .

To that end remark that, by definition,  $\mathcal{H}\Sigma^* = K^* = S^* - L^*$ , so that

$$\begin{aligned} \begin{pmatrix} \Sigma_X^n \\ 2\Sigma_{ZX}^n \end{pmatrix} - \mathcal{F}_{\Sigma_Z^n}(\hat{S}_\Omega - \hat{L}_{T'}) &= \begin{pmatrix} \Sigma_X^n \\ 2\Sigma_{ZX}^n \end{pmatrix} - \mathcal{F}_{\Sigma_Z^n}(\mathcal{H}\Sigma^* + \mathcal{A}(\Delta_S, \Delta_L)) \\ &= \begin{pmatrix} \Sigma_X^n \\ 2\Sigma_{ZX}^n \end{pmatrix} - \mathcal{F}_{\Sigma_Z^n}\mathcal{H}\Sigma^* - R_{K^*}(\mathcal{A}(\Delta_S, \Delta_L)) + \mathcal{I}_{\Sigma_Z^n}^*\mathcal{A}(\Delta_S, \Delta_L) \\ &\triangleq E_n - R_{K^*}(\mathcal{A}(\Delta_S, \Delta_L)) + \mathcal{I}_{\Sigma_Z^n}^*\mathcal{A}(\Delta_S, \Delta_L) \\ &= E_n - R_{K^*}(\mathcal{A}(\Delta_S, \Delta_L)) + \mathcal{I}_{\Sigma_Z^n}^*\mathcal{AP}_Y(\Delta_S, \Delta_L) + \mathcal{I}_{\Sigma_Z^n}^*\mathcal{C}_{T'}. \end{aligned}$$

The rest of the proof uses Corollary 2 and mirrors (Chandrasekaran 2011), Proposition B.4.2.  $\square$

## D.2 Variety constraints

We now turn to a variant of the problem in which both the rank and the sparsity pattern are enforced explicitly. This amounts to minimising the objective function within the following non-convex constraint set:

$$\mathcal{M} = \{(S, L) | S \in \Omega(S^*), \text{rank}(L) \leq \text{rank}(L^*),$$

$$\|\mathcal{P}_{T^\perp}(L - L^*)\|_2 \leq \frac{\xi(T)\lambda_n}{D\psi_X^{*2} \left(1 + 2\frac{\psi_Z}{\psi_X^*} (1 + \psi_X^* \phi_{ZX}^*)^2\right)},$$

$$g_\gamma(\mathcal{A}^\dagger \mathcal{I}_{\Sigma_Z}^* \mathcal{A}(S - S^*, L^* - L)) \leq 11\lambda_n\}.$$

We start by proving a preliminary result that bounds the spectral norm of the FIM evaluated at the nominal parameters. This simple bound will be used later in this section.

**Proposition 6.** Let  $K^* \triangleq \mathcal{H}(\Sigma^*)$  and recall that  $\mathcal{I}_{\Sigma_Z}^* \triangleq \mathcal{I}_{\Sigma_Z}(K^*)$ .

$$\|\mathcal{I}_{\Sigma_Z}^*\|_2 \leq \psi_X^{*2} \left( 1 + 2 \frac{\psi_Z}{\psi_X^*} (1 + \psi_X^* \phi_{ZX}^*)^2 \right).$$

*Proof.* We have

$$\mathcal{I}_{\Sigma_Z}^* = - \begin{pmatrix} K_X^{*-1} \otimes K_X^{*-1} & 0 \\ 0 & 0 \end{pmatrix} - 2K_X^{*-1} \otimes \begin{pmatrix} K_X^{*-1} K_{ZX}^{*T} \Sigma_Z^* K_{ZX}^* K_X^{*-1} & -K_X^{*-1} K_{ZX}^{*T} \Sigma_Z^* \\ \cdot & \Sigma_Z \end{pmatrix}.$$

Therefore,

$$\begin{aligned} \|\mathcal{I}_{\Sigma_Z}^*\|_2 &\leq \psi_X^2 + 2\psi_X^* \left\| \begin{pmatrix} K_X^{*-1} K_{ZX}^{*T} \Sigma_Z^* K_{ZX}^* K_X^{*-1} & -K_X^{*-1} K_{ZX}^{*T} \Sigma_Z^* \\ \cdot & \Sigma_Z \end{pmatrix} \right\|_2 \\ &= \psi_X^{*2} + 2\psi_X^* \left\| \begin{pmatrix} K_X^{*-1} K_{ZX}^{*T} \sqrt{\Sigma_Z^*} \\ \sqrt{\Sigma_Z^*} \end{pmatrix} \begin{pmatrix} K_X^{*-1} K_{ZX}^{*T} \sqrt{\Sigma_Z^*} \\ \sqrt{\Sigma_Z^*} \end{pmatrix}^T \right\|_2 \\ &\leq \psi_X^{*2} + 2\psi_X^* (\psi_X^* \phi_{ZX}^* \sqrt{\psi_Z} + \sqrt{\psi_Z})^2 \\ &\leq \psi_X^{*2} (1 + 2 \frac{\psi_Z}{\psi_X^*} (1 + \psi_X^* \phi_{ZX}^*)^2). \end{aligned}$$

□

We now have the following results. Remark that because we derived different bounds on the curvature of the rank variety and on the spectral norm of  $\mathcal{I}^*$ , the assumptions of these propositions differ from the one made in (Chandrasekaran 2011).

**Proposition 7.** (see (Chandrasekaran 2011), Proposition B.4.3)

Consider any  $(S, L) \in \mathcal{M}$  and let  $(\Delta_S, \Delta_L) = (S - S^*, L^* - L)$ . For  $\gamma$  in the range required for consistency and letting  $C_1 = \frac{48}{\alpha} + \frac{1}{\psi_X^{*2} \left( 1 + 2 \frac{\psi_Z}{\psi_X^*} (1 + \psi_X^* \phi_{ZX}^*)^2 \right)}$ , we have that  $g_\gamma(\Delta_S, \Delta_L) \leq C_1 \lambda_n$ .

**Corollary 3.** (see (Chandrasekaran 2011), Corollary B.4.1)

Consider any pair  $(S, L) \in \mathcal{M}$  and, as before, let  $\Delta_S = S - S^*$ ,  $\Delta_L = L^* - L$ . For  $C_1$  defined as in the previous proposition, further define the following constants:

- $C_2 = \left( \frac{24(2-\nu)}{\nu} \right) C_1^2 D \psi_X^{*2} \left( 1 + 2 \frac{\psi_Z}{\psi_X^*} (1 + \psi_X^* \phi_{ZX}^*)^2 \right);$
- $C_3 = C_1 + \frac{3\alpha C_1^2 (2-\nu)}{4(3-\nu)};$
- $\sigma = \text{smallest non-zero singular value of } L^*;$

- $C_4 = \max\{C_2, C_3\};$
- $C_5 = \frac{C_1\nu\alpha}{\beta(2-\nu)};$
- $T' = T(L)$  and  $\mathcal{C}_{T'} = \mathcal{P}_{T'\perp}(L^*)$ .

Assume that  $\sigma \geq \frac{C_4\lambda_n}{\xi(T)^2}$ , and suppose that the smallest entry in magnitude of  $S^*$  is greater than  $\frac{C_5\lambda_n}{\mu(\Omega)}$ .

Then we have:

1.  $L$  has rank equal to  $L^*$ ;
2.  $\text{sign}(S) = \text{sign}(S^*)$ ;
3.  $\|\mathcal{P}_{T^\perp}(\Delta_L)\|_2 \leq \frac{4\xi(T)\lambda_n}{73D\psi_X^{*2}\left(1+2\frac{\psi_Z}{\psi_X^*}(1+\psi_X^*\phi_{ZX}^*)^2\right)}.$
4.  $\rho(T, T') \leq \frac{8\xi(T)}{73}.$
5.  $g_\gamma(\mathcal{A}^\dagger \mathcal{I}_{\Sigma_Z}^* \mathcal{C}_{T'}) \leq \frac{\lambda_n\nu}{6(2-\nu)}.$
6.  $\|\mathcal{C}_{T'}\|_2 \leq \frac{16(3-\nu)\lambda_n}{3\alpha(2-\nu)}.$

### D.3 From variety to tangent space constraints

**Proposition 8.** (see, (Chandrasekaran 2011), Proposition B.4.4)

Let  $\gamma$  be in the correct range. Suppose that the minimum non-zero singular value of  $L^*$  is such  $\sigma \geq \frac{C_4\lambda_n}{\xi(T)^2}$ , and suppose that the smallest entry in magnitude of  $S^*$  is greater than  $\frac{C_5\lambda_n}{\mu(\Omega)}$ . Let  $g_\gamma(\mathcal{A}^\dagger E^n) \leq \frac{\lambda_n\nu}{6(2-\nu)}$ . Further suppose that

$$\lambda_n \leq \frac{3\alpha(2-\nu)}{16(3-\nu)} \min\left(\frac{1}{6\psi_X^*}, \frac{\phi_{ZX}^*}{4}, \frac{\alpha\xi(T)}{96D\psi_X^*\psi^2(1+\frac{\alpha}{6\beta})^2}\right).$$

Then we have  $(\hat{S}_\Omega, \hat{L}_{T_{\mathcal{M}}}) = (\hat{S}_{\mathcal{M}}, \hat{L}_{\mathcal{M}})$ .

**Corollary 4.** Under the assumptions of Proposition 8, we have that  $\text{rank}(\hat{L}_{T_{\mathcal{M}}}) = \text{rank}(L^*)$  and that  $T(\hat{L}_{T_{\mathcal{M}}}) = T_{\mathcal{M}}$ . Similarly,  $\text{sign}(\hat{S}_\Omega) = \text{sign}(S^*)$ .

#### D.4 From tangent space constraints to problem (A.1)

**Lemma 1.** (see, (Chandrasekaran 2011), Lemma B.4.1)

Let  $(\hat{S}_\Omega, \hat{L}_{T_M})$  be the solution to the variety constrained problem (A.2). Suppose that the assumptions of Proposition 8 hold. Further assume that

$$g_\gamma(\mathcal{A}^\dagger R_{K^*} \mathcal{A}(\Delta_S, \Delta_L)) \leq \frac{\lambda_n \nu}{6(2 - \nu)}.$$

Then  $(\hat{S}_\Omega, \hat{L}_{T_M})$  is also the unique optimum to problem (A.1).

#### D.5 Bounding the error terms

We start by recalling two useful results from the literature.

**Theorem 2.** ((Davidson and Szarek 2001), Theorem II.13)

Given  $n, p \in \mathbb{N}$ , with  $p \leq n$ , let  $\Gamma$  be  $p \times n$  matrix with i.i.d. Gaussian entries drawn from  $\mathcal{N}(0, 1/n)$ . Then the largest and smallest singular values of  $\Gamma$ ,  $\sigma_1(\Gamma)$  and  $\sigma_p(\Gamma)$ , are such that

$$\max \left( \mathbb{P} \left( \sigma_1(\Gamma) \geq 1 + \sqrt{\frac{p}{n}} + t \right), \mathbb{P} \left( \sigma_p(\Gamma) \leq 1 - \sqrt{\frac{p}{n}} - t \right) \right) \leq \exp \left( -\frac{nt^2}{2} \right),$$

for  $t > 0$ .

**Lemma 2.** ((Wytock and Kolter 2013), Lemma 1)

Let  $c_Z$  denote the maximum variance of the columns of the  $n \times m$  matrix  $Z$  (i.e the  $\|\cdot\|_{1 \rightarrow 2}$  operator norm). Let  $c_X^*$  denote the maximum diagonal entry of  $K_X^{*-1}$ . Then,

$$\mathbb{P} \left( \|2\Sigma_{ZX}^n + 2\Sigma_Z K_{ZX}^* K_X^{*-1}\|_\infty \geq \delta \right) \leq 2mp \exp \left( -\frac{n\delta^2}{8c_X^{*2} c_Z^2} \right).$$

Using these two results we can bound the spectral norm of the error term

$$E^n = \begin{pmatrix} \Sigma_X^n \\ 2\Sigma_{ZX}^n \end{pmatrix} - \mathcal{F}_{\Sigma_Z^n}(K^*).$$

We recall the notations defined in Proposition 4: we write  $\psi_Z \triangleq \|\Sigma_Z\|_2$ ,  $\psi_X^* \triangleq \|K_X^{*-1}\|_2$  and  $\phi_{ZX}^* \triangleq \|K_{ZX}^*\|_2$ .

**Proposition 9.** Let  $E^n$  be defined as above. Then for  $0 < \delta \leq 8\sqrt{2}\psi_X^*(1 + \psi_X^*\psi_Z\phi_{ZX}^{*2})$  and  $n \geq \frac{128p\psi_X^*(1+\psi_X^{*2}\psi_Z\phi_{ZX}^{*2})^2}{\delta^2}$ , we have that

$$\mathbb{P}(\|E^n\|_2 \geq \delta) \leq \max \left( 2 \exp \left( -\frac{n\delta^2}{128\psi_X^{*2}(1 + \psi_X^*\psi_Z\phi_{ZX}^{*2})^2} \right), mp \exp \left( -\frac{n\delta^2}{16\psi_Z^2\psi_X^{*2}} \right) \right).$$



*Proof.* First recall that  $\|E^n\|_2 = \left\| \begin{pmatrix} E_X^n \\ E_{ZX}^n \end{pmatrix} \right\|_2$  is such that

$$\max(\|E_X^n\|_2, \|E_{ZX}^n\|_2) \leq \|E^n\|_2 \leq \sqrt{2} \max(\|E_X^n\|_2, \|E_{ZX}^n\|_2).$$

Hence, we have that

$$\mathbb{P}(\|E^n\|_2 \geq \delta) \leq \max\left(\mathbb{P}\left(\sqrt{2}\|E_X^n\|_2 \geq \delta\right), \mathbb{P}\left(\sqrt{2}\|E_{ZX}^n\|_2 \geq \delta\right)\right).$$

We bound both terms separately.

We look at  $E_X^n$  first :  $E_X^n = \Sigma_X^n - (K_X^{*-1} + K_X^{*-1} K_{ZX}^{*T} \Sigma_Z K_{ZX}^* K_X^{*-1})$  which is positive definite and of the form  $\Sigma_X^n - A$  where each row of  $\Sigma_X^n$  is drawn from  $\mathcal{N}(0, A)$ . Moreover, we have that  $\|A\|_2 \leq \psi_X^*(1 + \psi_X^* \psi_Z \phi_{ZX}^{*2})$ , so that:

$$\begin{aligned} \mathbb{P}(\sqrt{2}\|E_X^n\|_2 \geq \delta) &\leq \mathbb{P}(\sqrt{2}\|A\|_2 \|(A^{-1/2} \Sigma_X^n A^{-1/2} - I)\|_2 \geq \delta) \\ &\leq \mathbb{P}\left(\|(A^{-1/2} \Sigma_X^n A^{-1/2} - I)\|_2 \geq \frac{\delta}{\sqrt{2}\psi_X^*(1 + \psi_X^* \psi_Z \phi_{ZX}^{*2})}\right) \end{aligned}$$

Now, remark that  $A^{-1/2} \Sigma_X^n A^{-1/2}$  is of the form  $\Gamma \Gamma^T$  with  $\Gamma$  a matrix of dimension  $p \times n$  satisfying the assumptions of Theorem 2. Therefore,

$$\begin{aligned} \mathbb{P}(\sqrt{2}\|E_X^n\|_2 \geq \delta) &\leq \mathbb{P}\left(s_1(\Gamma^2) \geq 1 + \frac{\delta}{\sqrt{2}\psi_X^*(1 + \psi_X^* \psi_Z \phi_{ZX}^{*2})}\right) + \mathbb{P}\left(s_p(\Gamma^2) \leq 1 - \frac{\delta}{\sqrt{2}\psi_X^*(1 + \psi_X^* \psi_Z \phi_{ZX}^{*2})}\right) \\ &\leq \mathbb{P}\left(s_1(\Gamma) \geq 1 + \frac{\delta}{4\sqrt{2}\psi_X^*(1 + \psi_X^* \psi_Z \phi_{ZX}^{*2})}\right) + \mathbb{P}\left(s_p(\Gamma) \leq 1 - \frac{\delta}{4\sqrt{2}\psi_X^*(1 + \psi_X^* \psi_Z \phi_{ZX}^{*2})}\right) \\ &\leq \mathbb{P}\left(s_1(\Gamma) \geq 1 + \sqrt{\frac{p}{n}} + \frac{\delta}{8\sqrt{2}\psi_X^*(1 + \psi_X^* \psi_Z \phi_{ZX}^{*2})}\right) + \\ &\quad \mathbb{P}\left(s_p(\Gamma) \leq 1 - \sqrt{\frac{p}{n}} - \frac{\delta}{8\sqrt{2}\psi_X^*(1 + \psi_X^* \psi_Z \phi_{ZX}^{*2})}\right) \\ &\leq 2 \exp\left(-\frac{n\delta^2}{128\psi_X^{*2}(1 + \psi_X^* \psi_Z \phi_{ZX}^{*2})^2}\right). \end{aligned}$$

We now turn to  $E_{ZX}^n$ .  $E_{ZX}^n = 2(\Sigma_{ZX}^n + \Sigma_Z K_{ZX}^* K_X^{*-1})$ . Recalling that the infinity norm is bounded above by the spectral norm, we conclude from Lemma 2 that

$$\mathbb{P}(\|E_{ZX}^n\|_2 \geq \delta) \leq mp \exp\left(-\frac{n\delta^2}{8\psi_Z^2 \psi_X^{*2}}\right).$$

□

Finally, we can prove a result (which is an extension of (Chandrasekaran 2011), Corollary B.4.3) that allows us to condition on the norm of the error terms as a function of the number of samples and the dimension of the problem,  $p$  and  $m$ .

**Corollary 5.** Let  $\delta_n = \max \left( \sqrt{\frac{256mp\psi_X^{*2}(1+\psi_X^*\psi_Z\phi_{ZX}^{*2})^2}{n}}, \sqrt{\frac{16mp\psi_Z^2\psi_X^{*2}}{n}} \right)$ . Then for  $n \geq 2mp$  and  $mp \geq 2$ ,

$$\mathbb{P}(\|E^n\|_2 \leq \delta_n) \geq 1 - mp \exp(-mp).$$

*Proof.*  $\delta_n$  is of the form

$$\max \left( \sqrt{k_1 \frac{mp}{n}}, \sqrt{k_2 \frac{mp}{n}} \right).$$

Start by noticing that if  $n \geq 2mp$ , then  $\sqrt{k_1 \frac{mp}{n}} \leq 8\sqrt{2}\psi_X^*(1 + \psi_X^*\psi_Z\phi_{ZX}^{*2})$  so that we can apply Proposition 9. We then proceed by disjunction elimination.

Assume  $k_1 \geq k_2$ . Then, by Proposition 9 we have:

$$\mathbb{P} \left( \|E^n\|_2 \geq \sqrt{k_1 \frac{mp}{n}} \right) \leq \max(2 \exp(-2mp), mp \exp(-k_1/k_2 mp))$$

Because  $k_1 \geq k_2$ ,  $mp \exp(-k_1/k_2 mp) \leq mp \exp(-mp)$ , so that

$$\max(2 \exp(-2mp), mp \exp(-k_1/k_2 mp)) \leq mp \exp(-mp),$$

as  $mp \geq 2$ .

If  $k_2 \geq k_1$ , then we have a similar proof, and the result follows.  $\square$

## D.6 Putting it all together

**Assumption 3.** (*Assumptions for Algebraic Consistency*)

1. Set

$$C_6 = \frac{\alpha\nu}{32(3-\nu)D} \min \left( \frac{1}{6\psi_X^*}, \frac{\phi_{ZX}^*}{4}, \frac{\alpha\nu}{384D(3-\nu)\psi_X^*\psi^2(1+\frac{\alpha}{6\beta})^2} \right)$$

and let  $n$  be such that

$$n \geq \frac{mp}{\xi(T)^4} \max \left( 2, \frac{256\psi_X^{*2}(1+\psi_X^*\psi_Z\phi_{ZX}^{*2})^2}{C_6^2}, \frac{16\psi_Z^2\psi_X^{*2}}{C_6^2} \right)$$

2. Set

$$\delta_n = \max \left( \sqrt{\frac{256mp\psi_X^{*2}(1+\psi_X^*\psi_Z\phi_{ZX}^{*2})^2}{n}}, \sqrt{\frac{16mp\psi_Z^2\psi_X^{*2}}{n}} \right)$$

and let

$$\lambda_n = \frac{6D\delta_n(2-\nu)}{\xi(T)\nu}$$

3. Let the minimum non-zero singular value of  $L^*$  be such that

$$\sigma \geq \frac{C_4 \lambda_n}{\xi(T)^2}.$$

4. Let the minimum magnitude nonzero entry  $\theta$  of  $S^*$  be such that

$$\theta \geq \frac{C_5 \lambda_n}{\mu(\Omega)}.$$

Under these assumptions Theorem (1) follows (see (Chandrasekaran, Parrilo and Willsky 2012), Section 5.5).

## REFERENCES

- Bach, F. (2008), “Consistency of trace norm minimization,” *Journal of Machine Learning Research*, 8, 1019–1048.
- Bowden, J., Davey Smith, G., and Burgess, S. (2015), “Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression.,” *International journal of epidemiology*, 44(2), 512–25.
- Boyd, S. (2011), “Alternating Direction Method of Multipliers,” *Proceedings of the 51st IEEE Conference on Decision and Control*, 3(1), 1–44.
- Candès, E. J., Li, X., Ma, Y., and Wright, J. (2011), “Robust Principal Component Analysis?,” *J. ACM*, 58(3), 11:1–11:37.
- Chandrasekaran, V. (2011), *Convex Optimization Methods for Graphs and Statistical Modeling*, PhD thesis, MIT.
- Chandrasekaran, V., Parrilo, P. A., and Willsky, A. S. (2012), “Latent variable graphical model selection via convex optimization,” *The Annals of Statistics*, 40(4), 1935–1967.
- Chandrasekaran, V., Recht, B., Parrilo, P. a., and Willsky, A. S. (2012), “The Convex Geometry of Linear Inverse Problems,” *Foundations of Computational Mathematics*, 12(6), 805–849.
- Chandrasekaran, V., Sanghavi, S., Parrilo, P. a., and Willsky, A. S. (2009), “Rank-Sparsity Incoherence for Matrix Decomposition,” *SIAM Journal on Optimization*, 21(2), 572–596.
- Chen, L., and Huang, J. (2014), “Sparse reduced-rank regression with covariance estimation,” *Statistics and Computing*, (0960-3174), 1–10.
- Combettes, P. L., and Pesquet, J.-C. (2009), “Proximal Splitting Methods in Signal Processing,” in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pp. 3–30.
- Davidson, K. R., and Szarek, S. J. (2001), “Handbook of the Geometry of Banach Spaces,” in *Handbook of the Geometry of Banach Spaces*, eds. W. Johnson, and J. Lindenstrauss, chapter 8, pp. 317–366.

- Friedman, J., Hastie, T., and Tibshirani, R. (2008), “Sparse inverse covariance estimation with the graphical lasso,” *Biostatistics*, 9(3), 432–441.
- Kang, H., Zhang, A., Cai, T. T., and Small, D. S. (n.d.), “Instrumental Variables Estimation With Some Invalid Instruments and its Application to Mendelian Randomization,” *Journal of the American Statistical Association*, .
- Koller, D., and Friedman, N. (2009), *Probabilistic Graphical Models: Principles and Techniques*, Vol. 2009 of *Adaptive Computation and Machine Learning* MIT Press.
- Lauritzen, S. (1996), *Graphical Models: Hardback: Steffen L. Lauritzen - Oxford University Press* Oxford University Press.
- Nesterov, Y. (2005), “Smooth minimization of non-smooth functions,” *Mathematical Programming*, 103(1), 127–152.
- O’Donoghue, B., and Candès, E. (2013), “Adaptive Restart for Accelerated Gradient Schemes,” *Foundations of Computational Mathematics*, (12), 1–18.
- Parikh, N., and Boyd, S. (2014), “Proximal Algorithms,” *Foundations and Trends in Optimization*, 1(3), 123–231.
- Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. (2011), “High-dimensional covariance estimation by minimizing l1-penalized log-determinant divergence,” *Electronic Journal of Statistics*, 5(January 2010), 935–980.
- Rothman, A. J., Levina, E., and Zhu, J. (2010), “Sparse Multivariate Regression With Covariance Estimation,” *Journal of Computational and Graphical Statistics*, 19(4), 947–962.
- Schmidt, M., Roux, N. L., and Bach, F. (2011), “Convergence Rates of Inexact Proximal-Gradient Methods for Convex Optimization,” *Advances in Neural Information Processing Systems 24*, abs/1109.2, 1–9.
- Sohn, K.-A., and Kim, S. (2012), Joint Estimation of Structured Sparsity and Output Structure in Multiple-Output Regression via Inverse-Covariance Regularization,, in *Conference on Artificial Intelligence and Statistics*.

- Stewart, G., and Sun, J.-g. (1990), *Matrix Perturbation Theory* Academic Press.
- Wainwright, M. J. (2009), “Sharp thresholds for high-dimensional and noisy sparsity recovery using  $\ell_1$ -constrained quadratic programming (Lasso),” *IEEE Transactions on Information Theory*, 55(5), 2183–2202.
- Wytock, M., and Kolter, J. Z. (2013), Sparse Gaussian conditional random fields: Algorithms, theory, and application to energy forecasting,, in *Proceedings of the 2013 International Conference on Machine Learning (ICML '13)*, eds. S. Dasgupta, and D. Mcallester, Vol. 28, JMLR.org, pp. 125–1273.
- Zhang, L., and Kim, S. (2014), “Learning gene networks under SNP perturbations using eQTL datasets.,” *PLoS computational biology*, 10(2), e1003420.